Author:
Alam, Justin Shumon

Title:
Radical evil, freedom and moral self-development in Kant's practical philosophy

Radical Evil, Freedom and Moral Self-Development in Kant's Practical Philosophy

by

Justin Shumon Alam

A dissertation submitted to the University of Bristol in accordance with the requirements for the degree of Doctor of Philosophy in the Faculty of Arts

School of Arts
January 2011

75,245 words

Abstract

Kant remains an important resource in moral philosophy but the absence of an adequate account of moral self-development constitutes a serious gap in his wider moral theory. This study therefore seeks to illuminate the process through which an agent could develop his moral character within a Kantian framework. Firstly, I reject two interpretations of Kant's account of rational agency each of which, if true, would in its own way render moral development impossible. I also outline the interpretation of Kantian rational agency which I take to be correct and which allows development. Kant thinks development should address our radical evil - an attitude to choice which rejects the demands of the moral law. However, there are tensions in the doctrine of evil which seem to preclude an evil agent's initiating his moral development. I adopt Seiriol Morgan's rational reconstruction of evil which addresses these difficulties. In Morgan's model, the will's freedom gives it overriding reason to choose morality, as this affords it its true freedom and it knows this. This means the will which chooses evil must wilfully accept a false conception of freedom - it must be self-deceived at the most fundamental level of reason-giving. However, self-deception is *prima facie* paradoxical. This is addressed by applying Jean-Paul Sartre's model of bad faith, an account which can dissolve the paradoxes. What emerges is a picture of evil as a mutually supporting complex of elements involving selfishness, self-conceit and a refusal to acknowledge its own misguided attitude. This is the opponent for morality. Development involves undoing this structure through consciousness of true freedom and pursuing the ends of development such as the purification of motives, whilst remaining vigilant against further deception. In this way, the free will can acquire a character apt to express its full freedom.

Dedication and Acknowledgements

I dedicate this study to my mother and father. I would not have been able to do it without your love and unfailing practical support. I also owe a great debt to my supervisor, Seiriol Morgan, since his work is foundational to this study. Moreover, I thank him for the time and effort he has put into discussing the project and the fruitful directions towards which he has gestured at decisive junctures. I also thank Jimmy Doyle for his support. I am very grateful to Jonathan Webber, who, both in conversation and in writing has actually managed to make Sartre clear. I thank those (too numerous to mention individually) who have listened to and engaged with me over the issues in this study. To my friends abroad - Selma, Funda, Dan, Ed and Steve - thanks for asking after me. I want to express to my friends here - Hannah, Chris and Anthony - and back home - Chris and Piers - my warmest appreciation for their being good friends. Finally, thank you to Henry and Sophie for all of their calls to their uncle to say hello.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidates own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED:  .....DATE: 4th July 2011.....

Contents

List of Abbreviations

A/B	*Critique of Pure Reason*
G	*Groundwork of the Metaphysics of Morals*
KpV	*Critique of Practical Reason*
MA	*Speculative Beginning of Human History*
MS	*The Metaphysics of Morals*
R	*Religion Within the Boundaries of Mere Reason*
VA	*Anthropology from a Pragmatic Point of View*
VE	*Lectures on Ethics*

In the case of the *Critique of Pure Reason*, references are made to the standard A and B pagination of the first and second editions. In the case of all the other works of Immanuel Kant referenced, the standard abbreviations (listed above) are used, along with volume:page number of the Prussian Academy edition of *Kants gesammelte Schriften*. In both cases, references are made in parentheses in the text.

Chapter 1

Introduction

Arguably, Kant's practical philosophy has yielded one of the most important, influential and enduring moral theories in the Western tradition. To many of us, the intuitive appeal of its basic claims is clear. To begin with, it provides a compelling answer to the question 'Why should I be moral?' by arguing that we *just are* moral beings because we are conscious of moral demands and that we are conscious of these simply because of the way we are constituted: as free and rational agents. Rather than being sensuously necessitated - that is, simply 'pushed around' by our desires - we are sensuously *affected* rational beings. A person has a will consisting in practical reason and is thus free to reflect on desires and decide whether to accept them as *reasons* for action. And whilst we may potentially take anything as a reason, *qua* rational beings we take ourselves to be obligated to reject any policy which cannot be universalized. We are thus moral agents in virtue of being free and rational agents who recognize the authority of a universalizability test.

Kant thinks our consciousness of the moral law shows us that we are free because the demands are *unconditional* (they make no reference to any particular desire issuing from the causally deterministic world) and we are also aware that we are *capable* of obeying them. They must come from us *qua* rational beings and not from our empirical nature. The notion that we are truly free from causal determination in the choices we make is perhaps another point which agrees with the first-personal experience many suppose they have as agents making choices. In standard cases at least, we do not simply find ourselves doing one thing, then doing another without our having had some say in the matter. From the practical standpoint, the standpoint of the agent, we have a genuine choice. Kant is able to make his agent responsible for her actions because she is the cause of them through her free choices and, for some at least, a moral theory which explains why a person can be held accountable is an appealing one.

We may also have the intuition that *persons* are to be respected in the sense that they should not be treated as we treat *things* and Kant's theory provides reasons why this should be so. The thought is that the moral law is of incomparable value since it commands a rational and free being to do that which is *supremely* rational and affords it its full freedom. Since the law inheres in persons, to violate them is to violate the law, the very source of supreme value in the world. The categorical imperative grounds a notion of rights which we have simply as persons. This prohibits action-types which are obviously wrong (such as rape

and torture) but it also provides compelling reasons why we may not, for example, show kindness in a way which does not respect a person's humanity. Also, even an individual whom we judge we have reason to hate or even despise is worthy of our respect because whatever he has done or failed to do, the potential to be good inherent in his reason demands it. In this way at least, Kant's moral philosophy is one of hope.

As a moral theory in which reason and universalizability are central, it offers a theoretical basis for other ideas that many would find congenial: a notion of fairness, of equal treatment of relevantly similar individuals, of not arbitrarily making oneself (or anyone else) a special case. And since a moral action is done for an overriding reason, it is the sort of thing for which we can offer a rational justification to our fellow rational beings. In addition, since an end which agrees with a supremely rational principle is one to which no rational being could reasonably object, a world in which all ends were like this would be one in which all ends harmonized. This is the world which the moral law demands we pursue in our choices. It is what Kant calls the Kingdom of Ends. I for one think that if one were looking for an ideal world towards which humankind can and should approximate, this is an excellent candidate, to say the least.

Whilst Kant's practical philosophy might be recommended by all of these features, the revival of virtue theory, beginning in the latter part of the twentieth century (perhaps with G.E.M. Anscombe's 'Modern Moral Philosophy')[1] brought with it various attacks on his account. One example was the call to do away with the notion of moral obligation which was seen as an outmoded concept linked to a religious ethical framework whose currency was already on the wane. This is a line of criticism which can perhaps be traced back to Friedrich Nietzsche in, for example, *On the Genealogy of Morals*.

A further point is that Kantian ethics excludes the emotions from having any kind of central role in moral agency, whereas, in virtue theory, having the right sort of emotion, in the right situation, in relation to the right people etc., is praiseworthy. There was also an objection to the notion that life can be divided into that portion to which obligation applies and that to which it does not. The problem was taken to be that there might be activities (such as the development of one's talents) which, one can agree, are not obligatory but are nonetheless activities for which we might suppose a person should be criticized for ignoring. This was advanced particularly by Bernard Williams in his book, *Ethics and the Limits of Philosophy*.[2]

---

[1] Reprinted in Crisp & Slote (Eds.), (1997).

[2] See Chapter 10 in Williams, (1993).

Another criticism put forward by Williams in that book[3] is that the sort of motivation involved in moral choices in Kant may lead to a diremption of the self since Kant merely assumes that practical deliberation is impartial and detached in the way that factual deliberation is. But since, as Williams believes, the former is 'first-personal, radically so', it 'involves an *I* that must be more intimately the *I* of my desires than this account [Kant's] allows,' (1993, p.67) attempts to adhere to Kantian morality may, in certain circumstances, call for one to abandon that which one cherishes most in life. Similar concerns were advanced by Michael Stocker in his 'The Schizophrenia of Modern Ethical Theories'.[4] Williams' charge of diremption relies on a rejection of Kant's entire account of rational agency and it is beyond the scope of this study to respond to it fully. However, in the process of pursuing its main concerns, there emerges the beginnings of a response to it[5] in Chapter 6.

The criticism most relevant to the present study was that of Kant's supposed over-emphasis on discrete acts and, correlatively, the neglect of *character*. This charge was levelled, for example, by Alasdair MacIntyre in *After Virtue*[6] but as recently as 1997, Roger Crisp and Michael Slote, editors of the collection of readings *Virtue Ethics* write in their Introduction,

> There is no doubt that modern ethics has indeed concentrated, in a legalistic fashion, upon rules concerning particular actions. But is it possible for utilitarians and Kantians to enlarge the focus of their own theories to incorporate agents' lives as a whole, their characters as well as or even instead of their actions? (1997, p.3)

In contrast, they say, a 'striking feature of virtue ethics is its focus on moral agents and their lives, rather than on discrete actions . . . construed in isolation from the notion of character, and the rules governing these actions.' (ibid.)

However, in actual fact, Kant *does* have a theory of character. But this was overlooked for many years, perhaps because attention was focussed on the works of the 1780's - *Groundwork of the Metaphysics of Morals*[7] and *Critique of Practical Reason*[8] - in which character receives little sustained treatment. These may have been the focus because it is in them that Kant attempts to carry out the crucial tasks of establishing that we are subject to morality and that we are free. In addition, since the categorical imperative plays a significant role in these works, they provide the basis for much (distracting) discussion of the intuitiveness or otherwise of the purported results of its universalizability test, i.e., the *acts* it

---

[3] See Chapter 4 in Williams, (1993).

[4] Reprinted in Crisp & Slote (Eds.), (1997).

[5] See Chapter 6, Section 1 of this study.

[6] (1981, p.219).

[7] From now on, I will refer to this work as the *Groundwork*.

[8] Hereafter, this will be referred to as the second *Critique*.

finds to be morally good, bad and indifferent. Thus, in focussing only on these works and the sort of discussions they encourage and not on those works containing Kant's account of character, he was perceived as mainly or even exclusively an act-theorist. However, from the early 1990's onwards, attention switched to the hitherto neglected works of the 1790's in which most of the sustained treatment of character is to be found (perhaps in response to the criticisms of the friends of virtue theory). These works are *Religion Within the Boundaries of Mere Reason*[9], *The Metaphysics of Morals* and *Anthropology from a Pragmatic Point of View*.[10]

There are many facets to Kant's account of character and many contributors to its interpretation and in some cases, reconstruction. Onora O'Neill, amongst others, has argued that maxims[11] - the subjective practical principles upon which we base our actions - are *general* principles upon which we act in various ways depending on the situation. Arguably, at least some maxims are of a high level of generality and we can see that this would make them truly elements of character rather than mere parochial, context-specific intentions.

Moreover, in his book *Kant's Theory of Freedom*, Henry Allison has provided an interpretation of the largely unfamiliar and in some ways difficult notion of *Gesinnung* (usually translated as 'disposition'), which can be described as an overarching policy, a supreme maxim or *meta-maxim*[12] which guides our choices of other lower order maxims. Christine Korsgaard also provided a clear account of *Gesinnung* and its connection with our other maxims in her essay 'Morality as Freedom'.[13] The idea of a maxim and of a supreme maxim together form a conception of a character which is *merely intelligible* and according to Kant's strictures, it would seem, unknowable. However, Allison argues that we can approximate an individual's maxims from her outward habitual behaviour. This can be said to constitute her empirical character.

Arguably, no account of character is complete without an account of moral character development. In the 1990's, work which clarified Kant's fairly complex taxonomy of duties[14] helped to highlight the fact that he was interested in this. A distinction was explicated

---

[9] I refer to this work as the *Religion*, from now on.

[10] From this point, I will refer to this work as the *Anthropology*.

[11] A notion that should be familiar even to those who had only read Kant's most widely-read work on moral philosophy: the *Groundwork*.

[12] The term 'meta-maxim' is Philip Quinn's and appears in his article 'In Adam's Fall, We Sinned All'. *Philosophical Topics 16* (1988): pp.89-118.

[13] Reprinted in her *Creating the Kingdom of Ends* (1996a).

[14] For which the main source is *The Metaphysics of Morals*.

between 'strict' duties - requirements to perform or omit a single action[15] - and 'wide' duties of virtue - each a requirement to set an end and develop oneself in respect of it: for example, the duty to be beneficent towards one's fellows. Chief amongst these is the duty to acquire virtue or moral strength - the capability Kant thinks we all have to do our duty from duty (which like physical strength is possessed by all but varies in degree between individuals). Other important ends include moral perfection - doing all of one's duties - and holiness - doing one's duty from duty.

Nancy Sherman's book, *Making a Necessity of Virtue* is an important contribution to the debate about Kant's vision of the virtuous individual and of the aims of moral development. Sherman is keen to highlight what she takes to be Kant's Stoic and Aristotelian influences within this vision. These are notably expressed, she argues, in the fact that the development project for Kant includes not just the end of strengthening the will itself but also the distinctly virtue ethical notion of character habituation. This involves cultivating those emotions which can potentially co-operate with duty so as to be responsive to the authority of reason. Moreover, she argues that Kant thinks that compassion can be cultivated with the aim of deploying it as a moral 'mode of attention' - a faculty through which to discern the morally salient features of situations. She also (more tentatively) suggests that it could function as a moral motive. Thus Sherman's Kantian agent aims to develop a kind of situational sensitivity and, possibly, rationally regulated emotional motives which are reminiscent of the virtues of an Aristotelian agent.

Sherman perhaps makes Kant's idea of the ends of development out to be more Aristotelian than they in fact were. However, a more serious shortcoming of her account is the continual supposition that these ends simply involve addressing wayward sensibility - the agent's desires. But by the time her book was published (1997) it had already been clear for several years in the literature (e.g., Allison, 1990) that Kant regards that attitude to choice which is wilfully in opposition to the law - *radical evil* - as the real target (and the active opponent) of moral development. Although she mentions evil in passing on one or two occasions, there is nothing remotely like a sustained discussion of it and its role in development and the terms 'evil', 'radical evil' and 'propensity to evil' do not appear at all in the index of her book. Since Sherman overlooks the underlying, characteristic reasons for immoral choice, she is not in a position to give an account of the challenges a would-be virtuous person faces in engaging in moral development.

In contrast to Sherman, there are those who acknowledge the relevance of evil to a Kantian moral development project. One such theorist is G. Felicitas Munzel in her book *Kant's*

---

[15] This is perhaps the sort of duty on which was based the caricature of Kant as a purely act-theorist.

Conception of Moral Character. The value of understanding this is that we appreciate that it is (at least primarily) the will itself and its attitude to choice (and not sensibility) which must be addressed. In addition, Munzel (amongst others) also recognizes that Kant supposes that evil requires self-deception since the moral law is supremely normative for a rational being and evil involves going against this in a way for which we are responsible. In Munzel's case, she argues that agents may deceive themselves into thinking that their bad or morally indifferent actions have moral worth when they merely have the outward appearance of moral ones. The recognition of the importance of evil and its supporting self-deception seems to allow her a way forward in giving an account of moral development: for example, she (quite reasonably) argues that truthfulness regarding the moral status of our actions is key to moral progress.

However, she ignores the fact that there are tensions in Kant's account of evil in the *Religion* which apparently threaten the preclusion of an evil agent's becoming a good one. Briefly, Kant believes one must be committed to either morality or evil but it is impossible to be committed to both. Evil is universal and inextirpable but we must (yet somehow) adopt morality to initiate moral development. It is not clear how she can proceed with an account of moral development without first addressing this. Also, Munzel does not acknowledge the apparent paradoxes one might associate with self-deception: firstly, this notion seems to suggest that we both believe something and believe its opposite (what I call the Belief Paradox) and secondly, that we both know about the deception and are ignorant of it (which I call the Deceiver Paradox). Until these paradoxes are dissolved, it would seem that evil cannot be facilitated in the way she envisages, nor is any solution to evil she proposes on the basis of such a picture of it worth putting forward. In addition, we might wonder why people might find themselves in the position of having to deceive themselves that they or their actions are good. Given that the law is overriding, what (pseudo)-reason do they suppose they have for doing something other than obeying the law in the first place? And how are they deceiving themselves that this is a genuine reason? (We will return to these issues shortly.)

A similar shortcoming to this last one is also to be found in Michael K. Green's paper, 'Kant and Moral Self-Deception'. He argues that in order to do wrong, agents must deceive themselves with regard to what is signified by their conscience. Again, I think this is correct as far as it goes. But it does not get to the heart of the matter: it is not clear why these agents think they have a reason for choosing evil in the first place and how they have managed to sell themselves this 'reason', i.e., how and why some are even in the business of trying to do things like denying the meaning of their pangs of conscience. Lawrence Pasternack's discussion in his paper 'Can Self-Deception Explain *Akrasia* in Kant's Theory of

Moral Agency?' *is* of a self-deception regarding such ultimate reason-giving (the overriding authority of the moral law versus the allure of self-love). But he quickly concludes that if such self-deception is itself evil and must be done intentionally, then such a maxim of self-deception would itself seem to require self-deception and we have a regress. He concludes that we cannot see how such self-deception regarding the acceptability of evil is possible in Kant.

The position, then, is this: Kant's practical philosophy, when properly understood, has a great deal to offer to wider ethical discourse and it is all the richer for having, as has been discovered more recently, a theory of character. However, it would constitute a serious gap in this latter theory if it contained no clear account of moral self-development.[16] Kant leaves himself open to the accusation of such an omission since whilst he tells us that there are duties of moral development (such as the acquisition of virtue), he gives the agent very little guidance on exactly *how* these are to be pursued. The agent's initial response might be to tackle radical evil since we know that Kant takes this to be the real opponent of morality (rather than sensibility). However, there are tensions in the account of radical evil, as it stands, and these seem to threaten the preclusion of the radical change of heart which Kant himself takes to be necessary to initiate moral development. If this can be addressed, there still remains the problem that evil requires self-deception at the most fundamental level. It is needed to make it seem as though evil has a degree of normativity to rival that of morality (whereas in reality it has no normativity in comparison). The problem is that (any) self-deception seems paradoxical. But if we can provide an account which dissolves its paradoxes and shows how it sustains evil, we might thereby acquire a clearer understanding of the structure of this opponent of morality and be in a better position to see how the agent might tackle it - i.e., how she might pursue a programme of moral self-development. I take it that these are the things I manage to do in the present study to move our understanding of Kant's account of character forward. I would now like to explain how I intend to do this.

Before proceeding with the project proper, some preliminary work is necessary. Thus, in Chapter 2 I argue against two conceptions of Kant's theory of rational agency, one of which would make evil impossible and the second of which would make moral development impossible. These obviously threaten a project that seeks to explicate moral development that targets evil. What I take to be the correct conception (the Incorporation Thesis) allows

---

[16] Kant does provide sustained accounts of what he takes to be good practice in the moral education of the young (i.e., other-person development), in the Doctrine of Method of the second *Critique*, in the Doctrine of Method of *The Metaphysics of Morals* and in *On Pedagogy*. However, my interest, in the present study, is primarily in the area of moral *self*-development, an area which receives no detailed or sustained treatment in the Kant corpus. It should also be noted I have taken those approaches to development which can be found in the *Critique of the Power of Judgment* (such as the contemplation of the sublime) to be beyond the scope of this study.

evil and moral development. I also address the notion of the 'timeless' choice of a supreme maxim by adopting the view that as a commitment (to good or evil), such a choice occurs at no particular time rather than being literally timeless.

In Chapter 3, I begin to outline the structure of evil with a view to understanding how it may be overcome or at least resisted. As I mentioned above, we find that there are tensions in Kant's account of radical evil in the *Religion* which if left unresolved, seem to preclude an evil agent from becoming a good one. I adopt Seiriol Morgan's rational reconstruction of evil which overcomes this and other difficulties. This model takes it that since a free will must regard a choice which affirms its freedom as maximally normative for it, the will must *misconstrue* freedom in order to choose evil. According to this conception of freedom, doing whatever one wants is freedom *simpliciter* (in spite of what others want or are entitled to demand). And since a policy of evil allows the affirmation of this conception, taking the latter as freedom makes evil seem choice-worthy. We now have a better understanding of the nature of the reason which agents think they have to choose evil.

However, since this misconstrual of freedom must be one for which the will is *responsible*, we are led to the supposition in Chapter 4 that the adoption of an evil supreme maxim requires that the will be self-deceived with regard to its conception of freedom. I deploy Jean-Paul Sartre's notion of bad faith to dissolve the paradoxes of self-deception mentioned above. The self-deception envisaged here is one associated with the most fundamental level of reason-giving (i.e., freedom as a reason) and it is with this account in particular that the present study begins to take us into new territory in Kant scholarship. In applying the notion of Sartrean bad faith to the Kantian will which chooses evil it becomes necessary to demonstrate that this will can do all of the things that Sartre's self-deception story says a self-deceiver must do to carry out the deception. A further challenge is to illuminate how the will can do something evil such as deceive itself into accepting an overarching policy of evil without its *already being evil*. This is similar to Pasternack's worry and the one which led him to admit defeat. The solution seems to be that the (many) elements of self-deception and the acceptance of evil are all 'equiprimordial'. The final challenge is to illuminate how the process of self-deception can be combined with the notion of a choice of supreme maxim that does not occur at any particular time.

Kant's exposition of the notion of *self-conceit* in the second *Critique* leaves it unclear whether it is part of the sensibility of the agent rather than a part of the orientation of his will and if it is the latter, whether it represents a second and more virulent source of evil than self-love. In Chapter 5, I argue that it is part of the orientation of the will but rather than a distinct source of evil it is yet another facet (along with the elements of self-deception regarding freedom) of

the evil attitude to choice. One final unclarity concerns the possibility that Kant takes all self-conceit to be quasi-moral. I argue that 'moral' self-conceit is one possible expression of self-conceit *simpliciter*.

Chapters 3-5 establish that to have an evil will is to have an attitude to choice consisting an in interlocking structure of mutually supporting elements of evil itself, a deliberate misconstrual of freedom and an inflated sense of self-worth. This is a new, more complex conception of the so-called *Denkungsart* or way of thinking. Its revelation allows us to see more clearly the true recalcitrance of the 'dream' from which the agent must 'wake' in order to adopt the moral supreme maxim and begin on the path of moral development. Exploring how this is possible is the initial challenge of Chapter 6. The key to a solution is the consciousness of *true* freedom (autonomy) but it is hard to see how the agent is to become conscious of it given her evil way of thinking. I argue that an initial good action, insignificant in itself, can initiate her awakening to freedom and morality.

In lifting self-deception and becoming aware of autonomy as true freedom, the agent *resolves* to be good. But, for Kant, adopting morality also requires a *commitment* to it consisting in moral progress. However, it is unclear whether this development can take all of the forms Kant thinks it does since some duties of development at first seem to clash with one another and this must be resolved before we can proceed. Also, there are elements of development which require moral self-knowledge for their pursuit (for example, the duty to purify one's motives) and it seems this would require that an agent's maxims are accessible to her. However, (direct) access to maxims is thought to be impossible. It therefore becomes necessary to find an indirect method through which an agent can glean some idea of her interests and motives. Although the reliability of the methods examined is limited by the possibility of self-deception, they seem sufficiently reliable to allow the pursuit of those elements of moral self-development which rely on self-knowledge.

Finally, I examine an interpretation of autonomy and virtue as forms of inner freedom. This is the freedom to be moved by respect for the law without temptation. Autonomy is said to be the sheer capacity for such freedom, virtue an enhanced capability for it. Since it also emerges that all elements of moral development require virtue in one way or another, it turns out that both the *initiation* of the process in the consciousness of autonomy as true freedom and the *affirmation* of that process (i.e., progress) involve consciousness of inner freedom - in the first instance as sheer capacity and in the second, as enhanced capability. We saw earlier that the evil will freely imprisons itself through a deliberate misconstrual of freedom. We now see that it is through greater self-consciousness *qua* freedom that the will may also free itself. Morgan's rational reconstruction of evil highlights the centrality of that notion in

Kant's practical thought. If this suggests to modern Kantians that the resource of the practical philosophy is a more pessimistic one than previously supposed, then it is my hope that providing a way forward in moral self-development may mitigate this somewhat.

Chapter 2

Kant's models of freedom and rational agency

Kant takes it that we are beings subject to morality because we are free. One way to understand why is to examine a famous argument which he presents at the beginning of Section III of the *Groundwork*. There, he gives a definition of freedom which he says is negative: '*freedom* would be that property of such a causality [the will] that it can be efficient independently of alien causes determining it' (G 4:446). He says that from this negative conception springs a positive one: the will is a kind of causality and so subject to a law (on pain of being 'an absurdity') (G 4:447). However, the law of a free will cannot be the law of natural necessity (which governs nature) but instead must be its own law and so, Kant says, autonomy or 'being a law to itself' (ibid.) is a property of a free will. Kant thinks that to be a law to itself, a free will must be subject to the principle 'act on no other maxim than that which can also have as object itself as universal law.' (ibid.) Since this is the formula of the categorical imperative - the supreme principle of morality - he concludes that 'a free will and a will under moral laws are one and the same.' (ibid.)

This is (one version of) the argument which Henry Allison calls the Reciprocity Thesis.[17] It has this name because its conclusion is that morality and freedom reciprocally imply one another. Kant uses it in the *Groundwork* to deduce morality from freedom. Roughly speaking,[18] he argues there that when we adopt the practical standpoint - the standpoint of a rational agent - we are conscious of ourselves as rational and *free* beings. According to the Reciprocity Thesis, this means we are subject to morality. In the second *Critique* Kant takes it as a 'Fact of Reason' that we are simply conscious of moral demands when we enter into practical deliberation and that because these demands make no reference to interests arising from sensibility we recognize them as the demands of reason, that is, unconditional, moral demands. Again, given the Reciprocity Thesis, we must conclude that we are free from the practical point of view, he argues.

However, if we were not free beings, then we would not be subject to morality. And, furthermore, if we were not subject to morality, it would not be possible for us to be evil because to be evil is to violate the requirements of morality and one cannot violate a law to

[17] See chapter 11 in Allision (1990).

[18] The deductions of morality in the *Groundwork* and of freedom in the second *Critique* are fairly complex, unclear in many places and open to interpretation and it would not serve my present purpose to enter into a lengthy explanation of them. However, there is a summary of Kant's *Groundwork* III strategy for the deduction of morality (according to Seiriol Morgan) in Chapter 3, Sub-section 2.1 of this study.

which one is not subject. There is a reading of Kant's models of rational agency and freedom, perhaps most famously advocated by Henry Sidgwick, according to which, Kant posits a form of freedom in which we are only free when we act morally and unfree in non-moral action. I would like to begin by dispelling this reading since it would render pointless a project such as mine which seeks to illuminate moral self-development, which, as we saw in Chapter 1, takes evil as the opponent. Moreover, Sidgwick's reading would actually cause the collapse of Kant's entire moral project since both morality and evil depend on an agent's being free and responsible whatever he chooses. I will also argue against an interpretation of Kant's account of rational agency put forward by Allen Wood which makes our intelligible character out to be some timeless noumenal causality. This is another reading which would kill the present project at birth since (for one thing) it obviously precludes any notion of *progress*. It also precludes a change in fundamental moral orientation which Kant takes to be necessary for the initiation of development.

Instead, I will advocate the interpretation of rational agency propounded by Allison in his book *Kant's Theory of Freedom*: the Incorporation Thesis. In addition to being a correct reading of Kant, and being intimately related to the important doctrines of *Gesinnung* and the hierarchy of maxims (also explained in this chapter), it has the virtue of enabling the present study rather than rendering it pointless (as it would be if either evil or moral progress were impossible). Since it is also an account of how maxims are adopted and since these subjective practical principles underlie a very large proportion of anything (I can imagine) one might want to say about the practical philosophy, including what I wish to say about it, it seems important to clarify and defend it before dealing with any other matter. We will see its importance early on in this study since it is foundational to a rational reconstruction of evil, introduced in the sequel.

1. Rejection of Sidgwick's *reductio*

As mentioned above, Sidgwick's interpretation of Kant, it seems, ascribes to him a view of freedom in which agents are causally determined by nature *in non-moral action*.[19] That Sidgwick's reading is a misinterpretation of Kant's view of freedom (and is shared by others),[20] is reason enough for any sympathetic student of the practical philosophy to try to

---

[19] Albeit, it sees moral actions as free.

[20] In addition to Sidgwick, this view is expressed, for example, by Bernard Williams (1993, p.64) and Robert Paul Wolff (1986, p.211). Henry Allision claims this is a view held by many and can be traced back to Hegel (Allison, 1990, p.2). Whilst for the most part, Lewis White Beck espouses the view in *A Commentary on Kant's Critique of Practical Reason* that the will can never be unfree (1960, pp.203-5), he also says (1960, p.203) that the *Willkür* can be an *arbitrium brutum* - a claim which seemingly contradicts the former one.

correct it (or to add his voice to those who have already tried to do so).[21] However, as I mentioned above, I have an additional motive to do this since it would render my project pointless. Sidgwick's failure to interpret Kant's view of rational agency results from a failure to understand his view of freedom. He thinks Kant uses the word 'freedom' in two different ways. According to the first model, which Sidgwick calls Rational Freedom, 'Freedom = Rationality, so that a man is free in proportion as he acts in accordance with Reason.' (Sidgwick, 1962, p.511).[22] It seems that according to this model, an agent might either submit herself to brute necessitation by her impulses or act according to a principle of reason. Apparently, we are free when we follow reason because we take our 'real selves', as it were, to be our rational selves (or at least that part of us which responds to reason), so rule by reason is self-rule, hence an expression of freedom. Conversely, when impulse is necessitating us, it is 'mastering us' (ibid.), so we are, in that instance, unfree. The second kind of freedom Sidgwick ascribes to Kant is called Moral Freedom and is the freedom of choice an agent has between good and evil actions. Sidgwick says that in Kant, this type of freedom is manifested just as much in a choice of evil as it is in one of good and that Kant means this kind of freedom whenever he must connect freedom with moral responsibility or imputation.

Sidgwick is right to argue that *if* a theorist were to posit both of these conceptions of freedom together, it would indeed get him into at least two difficulties. Firstly, he says that these two freedoms are incompatible in the following way: 'if we say that a man is a free agent in proportion as he acts rationally, we cannot also say, in the same sense of the term, that it is by his free choice that he acts irrationally when he does so act.' (ibid.) It seems that if free action is rational action, then irrational action cannot be free. In addition to saying that Rational Freedom is that conception in which freedom is equated with rationality, Sidgwick also says that Rational Freedom is 'the Freedom that we realise in proportion as we do *right*' (ibid., p.512; emphasis added). So, it seems that according to Rational Freedom, only in acting rightly do we act (fully) rationally and only in such actions is freedom (fully) expressed. Recall that Sidgwick claims Kant also posits Moral Freedom according to which we act freely regardless of whether we choose good or evil. The alleged problem for Kant is that from all of this, we get the contradiction that it is both the case and not the case that freedom is expressed exclusively in right (i.e. rational) action.

The second problem which Sidgwick points out is another contradiction and is a consequence of the first. Since according to Rational Freedom, rational (right) acts are the only free acts,

---

[21] For example, Allen Wood (1984, p.80).

[22] The Appendix to *The Methods of Ethics* from which this quotation is taken is a reprint with some omissions of an article which originally appeared in *Mind*, 1888, Vol XIII, No. 51. The statement quoted here is one of those omissions.

they are the only imputable ones. This goes against the doctrine of Moral Freedom which says that all acts - good or evil - are imputable (because it says both are free). Obviously, any self-respecting theory ought not to contain self-contradictory concepts. In addition to the problems Sidgwick indicates, there is another: in contradicting the notion that all acts are imputable, Rational Freedom contradicts a major desideratum of the practical philosophy. The conception of freedom which results in our acts being estimable when right and excusable when wrong is anathema to such a philosophy.

The question of whether these criticisms are fair obviously depends upon whether Kant actually posits these two conceptions of freedom. I think Kant does posit what Sidgwick calls Moral Freedom and it is the freedom associated with what Kant and his modern commentators would call the will's power of choice or *Willkür*.[23] As we shall see later in this chapter, when we examine the doctrine of *Gesinnung* and the hierarchy of maxims, this power of choice is negatively free - in other words, it in no wise involves interference from external causal determination, regardless of whether it is exercised in adopting a moral, permissible or immoral maxim of action. For these reasons - as I hope to make clearer as we progress - acting on a maxim of any moral status is imputable to the agent in Kant. So, Sidgwick is right to ascribe the positing of 'Moral Freedom' to him because this is merely the ascription of the positing of a free power of choice to him.

We must now determine whether Kant also posits Rational Freedom alongside Moral Freedom. Much of what Sidgwick says about the former suggests that he is talking about autonomy. As I understand it, in Kant's thought, the term 'autonomy' is used in two related ways: firstly, to refer to a *property of the will*, namely that *capacity* of the will to determine itself according to its own laws - i.e. the moral law; and secondly, to refer to any instance of such determination, i.e. to instances of willing in which the aforementioned property is expressed. The term 'positive freedom' can be substituted for both of these uses of the term 'autonomy'.

The evidence which points to Sidgwick's having autonomy in mind when discussing Rational Freedom is that he quotes passages from Kant in which the latter is clearly discussing autonomy. For example, in order to demonstrate that Kant posits Rational Freedom, he

---

[23] I subscribe to the interpretation of the Kantian will for which Allison (1990) argues in Chapter 7 of *Kant's Theory of Freedom*. This interpretation is as follows: the combined will or *Wille* is divided into the power of choice or *Willkür* and the legislative will which (slightly confusingly) Kant also refers to as *Wille*. Throughout this study, unless otherwise specified the term 'will' refers to the power of choice or *Willkür*. Although it would be an excessive divergence from the main point of this chapter to justify Allison's reading here, one virtue of it is that it allows a clear understanding of autonomy as the will giving or being the law to itself: according to this reading, the legislative will gives or is the law to the power of choice.

quotes Kant as saying ' "freedom, whose causality can be determined only by the law, consists just in this, that it restricts all inclinations by condition of obedience to pure law." ' (KpV 5:78; quoted in Sidgwick, 1962, p.514) This is a clear reference to the property of autonomy. A few lines later, Sidgwick, in ascribing Rational Freedom to Kant, also quotes from the argument at the beginning of *Groundwork* III (G 4:446-447), in which Kant claims that a free will must be subject to its own laws (rather than those of nature) - i.e., the argument in which Kant establishes that a free will must have the property of autonomy. I said earlier that Kant posits negative freedom (what Sidgwick calls Moral Freedom); since Kant obviously posits autonomy, (which might be what Sidgwick calls Rational Freedom) it might at first seem that the problems of positing both Moral and Rational Freedom stand.

In his article, 'The Value of Humanity and Kant's Conception of Evil', Matthew Caswell attempts to answer this by pointing out that ' "positive freedom" entails only that an agent is subject to moral law, not that he or she acts in accordance with it' (Caswell, 2006b, p.646). Even in cases of heteronomous willing, the agent is still autonomous (positively free) in the sense that he is still the sort of agent for whom autonomous willing is possible; autonomy is here understood as a property of the will (and Caswell is excluding the second of the two uses of 'autonomy' I outlined above - i.e. an *instance* of autonomous willing). So, according to Caswell's understanding of autonomy (that it is a property only), no matter what the agent does, he is still free (i.e. his will still has that property), so it seems the putative problems outlined earlier do not arise. However, whilst it is true that autonomy is always a property of the will regardless of what it wills, this does not help to answer Sidgwick's criticism, which is, I think, that it is possible for an agent to fail to will rationally and so freely *on a given occasion*. And if we can say that this is a failure to will autonomously on a given occasion, then it seems to be a failure to be free on such occasions (which seemingly contradicts Moral Freedom since, according to this, all willing is free). So Caswell's appeal to the persistence of the *property* of autonomy even where one is willing heteronomously will not answer Sidgwick's criticism.

The way to answer Sidgwick is to point out that it is indeed possible to fail to act autonomously - i.e. to fail to act with positive freedom - but that is merely a failure to act according to a maxim which agrees with a supremely rational principle (the categorical imperative). However, this has got nothing to do with a failure to *choose* freely; that is to say, a failure to be (negatively) free from causal determination. However, it is clear in the following quotation that Sidgwick mistakenly supposes that Rational Freedom *in itself* includes an idea of free choice.[24] He claims that Kant deploys Rational Freedom

---

[24] I would say that autonomous willing does indeed always *involve* an act of free choice but that is only because all willing always does; it is not peculiar to autonomous willing.

> when he seeks to exhibit the independence of Reason *in influencing choice*, then in many though not all his statements he explicitly identifies Freedom with this independence of Reason, and thus clearly implies the proposition that a man is free in proportion as he acts rationally. (Sidgwick, 1962, p.513; emphasis added)

So for Sidgwick, a failure to express Rational Freedom is not just a failure to act according to the moral law, *it is also a failure to choose freely* and it is in this second failure that the conflict with Moral Freedom consists since Moral Freedom says all choice is free. However, a failure to express autonomy in no way involves a failure to *choose* freely - that is to say a failure to be free from causal determination - and crucially such involvement would be required in order for Kant's autonomy to conflict with his conception of freedom of choice in the way that Rational and Moral Freedom conflict. Recall that according to Sidgwick himself, we would have to be talking 'in the same sense of the term ['freedom']' (ibid., p.511) in order for his criticism to go through and since we are not, it does not. In short, whilst Kant does posit a negatively free power of choice (Sidgwick's Moral Freedom), he posits no conception of freedom (like Rational Freedom) which would conflict with it.

Although we can now see that the cause of Sidgwick's confusion is that he takes Kant to be positing a form of freedom such that a failure to act according to the moral law is also a failure to choose with negative freedom, it is unclear whether Sidgwick regards a failure to be Rationally Free as (*a*) consisting simply in being causally determined or (*b*) freely *allowing* causal determination to take over. Evidence for interpretation (*a*) is that he says, 'choice in such actions is determined not "freely" but "mechanically", by "physical" and "empirical" springs of action.' (ibid., p.515). However, he says elsewhere that an agent 'becomes subject to physical causation, to laws of a brute outer world' when he 'wrongly *allows* his actions to be determined by empirical or sensible stimuli' (ibid., p.516; emphasis added) and this suggests interpretation (*b*).

Although Sidgwick does not realise it, *if* Kant had posited either interpretation of Rational Freedom, it would have caused the latter serious problems in addition to those outlined at the beginning of this chapter. To illustrate the difficulty associated with interpretation (*a*), let us imagine an agent who is faced with some immoral temptation and with the option of doing the right thing. If he ends up acting immorally, it is hard to see how we can say he was causally determined to do so when the other 'direction' in which his behaviour could have gone involved free choice: if there were a free choice of a rational (i.e. moral) principle available to the agent, then in order not to do it, it must be refused. But refusing to adopt that principle is as much an act of freedom as adopting it would have been had he done so. This means doing the wrong thing cannot simply be being causally determined; it must involve some prior free refusal to do the right thing. In short, it is *impossible* to be causally

determined to do one thing if there is an alternative path which, if taken, is chosen freely. On interpretation (*a*), Sidgwick, in ascribing Rational Freedom to Kant, is ascribing an incoherent conception of freedom to him.

Interpretation (*b*) says that the agent who fails to choose the rational principle freely *allows* a natural impulse causally to determine his actions. If this is what Sidgwick envisages Rational Freedom to be, it would give rise to a different problem had Kant posited it. This conception must involve the claim that the individual freely *abdicates* as a rational chooser and actively allows temptation to take over the reins of his 'agency'. It allows for a kind of relative causal determinism in heteronomy; i.e. a picture of heteronomy consisting in a free choice to allow natural necessity to 'take over' and determine a series of one's actions. Whilst the initial abdication must be the result of a free choice and hence be imputable, it is, nevertheless, difficult to see how the rational power of choice could be returned to the agent. It would be like a free choice to euthanize one's freedom. Obviously, this would be a most unwanted result for Kant. Whether we interpret Sidgwick's notion of Rational Freedom as like (*a*), which is incoherent or (*b*), which involves the renunciation of one's agency, we will see in the next section that Kant's actual view of rational agency and the type of freedom associated with it avoid both of these problems.

2. The Incorporation Thesis

Negative freedom is a necessary condition of autonomous action: one cannot act on one's own law without first being free from necessitation by the other sort of law: natural necessity. It may be that because there is always negative freedom where there is autonomy, Sidgwick wrongly supposed that there must be an absence of negative freedom where there is no autonomy in the agent's action. Another point which may have led him to believe this is that an agent who fails to act morally does indeed act *as though* he is unfree by *freely* choosing to follow the path of something that is unfree. Even though in non-moral action the pathways followed are the same as those of a being determined by nature, it is still the agent, through the free choices of his will, that is thought to be responsible for his actions and not natural causation. And it is because of this that all of an agent's actions are imputable to him. However, in order fully to understand the relationship between a thoroughly free elective will and the empirical desires which are involved in determining it, we must now turn to the Incorporation Thesis.

That both versions (*a*) and (*b*) of Rational Freedom are wrong; that - contra (*a*) - there is no causal determination in non-moral action[25] in the practical philosophy and that - contra (*b*) - the abdication of a rational will is impossible are both brought out very clearly by Christine Korsgaard in the chapter entitled 'Analysis of obligation' in her book *Creating the Kingdom of Ends*. She says that for Kant, in acting under the idea of freedom, an agent must be taking herself to be acting 'on a principle which she must regard as voluntarily adopted' (Korsgaard, 1996a, p.57) and that agents as such cannot view themselves as 'simply impelled' into their actions. Korsgaard underlines how thoroughgoing Kantian freedom of choice is when she makes the further point that in the practical philosophy, when an agent is faced with an especially strong inclination and chooses to indulge it, she is indeed doing just that - *choosing* it. Its strength does not overwhelm, override or by-pass the power of choice but instead, when we say an agent satisfied a desire because it was a strong one, we mean that she took its strength to be a reason for choosing it.

2.1 Practical spontaneity[26] in the *Critique of Pure Reason*[27]

In making this point, Korsgaard is alluding to the 'Incorporation Thesis'. This is the name Henry Allison gives to his interpretation of Kant's model of rational agency whose development he traces in his book *Kant's Theory of Freedom*. Allison's work attempts (successfully, I think) to show that as early as the first *Critique*, Kant believed that in prudential (as well as moral) action, agents are not causally determined by their desires. In order to see why this is so, we must first note that (according to Allison) in the first *Critique*, Kant regards acting on the basis of imperatives as the basis of (if not actually identical to) rational agency. On this view, reason forms an imperative which serves as a touchstone, an objective rule, which the understanding may employ in order spontaneously to judge whether the 'course of action is "right" or "permissible" '[28] (Allison, 1990, p.39).

In addition, Allison points out that in the first *Critique* account, there is in all agency another act of spontaneous judgement 'through which [in the case of prudential actions] the

---

[25] I.e., that freedom is always expressed.

[26] My understanding of spontaneity is that it is the capacity of practical reason (and the understanding) to choose or take something as appropriate (and thereby to be a kind of cause) independently of the causal series of nature.

[27] From this point on I refer to this work as the first *Critique*.

[28] As I understand it, this is the sort of deliberation, which if characterized in terms of the ideas of the *Groundwork*, could be described as an agent's comparison of her proposed maxim - her subjective practical principle - with an objective practical rule. Such a rule is a rational standard by which her maxim - her policy - can be judged. If she has already willed an end, *p*, the relevant sort of objective rule is a hypothetical imperative. She can refer to it to decide whether her idea of what to do to achieve *p* will achieve it. In contrast, a categorical imperative is an objective rule which says what she ought to will no matter what her ends are. Again, she can use this rule to decide whether her idea of what to do is what she ought to do but with this rule, the 'ought' is unconditional.

inclination or desire is deemed or *taken as* an appropriate basis of action.' (ibid.; emphasis added) Since the activity of reason in forming an objective rule and the activities of the understanding in (*a*) judging a course of action to be appropriate according to that rule and (*b*) its activity of endorsing the relevant desire in the first place are all spontaneous, they are not part of the causal order. (It is important that what motivates the acceptance of the desire is itself not part of the causal order since we would then have causal determinism again.) So, in Kant, the desire does not act upon an agent in some mindless fashion like gravity might and instead, in order to act at all, the agent must 'adopt a model of deliberative rationality' (ibid., p.38) which involves ineliminably the consciousness of a 'moment of spontaneity', as Allison puts it, in which the desire is *taken as* an appropriate basis of action. This he thinks is illustrated in the first *Critique* (A534/B562), where Kant 'remarks that in considering a free act we are constrained to consider its "cause" as "not . . . so determining that it excludes the causality of our will"' (Allison, 1990, p.39). The point here is that although an inclination seen from the standpoint of empirical character might be a sufficient cause of action, from the practical standpoint, what is required to determine the will is a 'complement of sufficiency' in the form of the spontaneity of the agent in a merely intelligible act of *taking* the inclination *as* an appropriate basis for action.

Allison regards this act of taking a desire as appropriate as constituting what is known as the intelligible character of rational agency. This is the character of rational agency which cannot be experienced (because one cannot observe oneself deciding) but which can be thought. His analysis of the first *Critique* account of rational agency makes it clear that even at this stage of Kant's thought, agents are not causally determined to act but instead are conscious of endorsing a desire as appropriate or permissible. Given that they are self-consciously endorsed in this way, action on the basis of them is imputable to the agent. Kant's account of the model of deliberative rationality which agents must adopt in order to act is developed in the *Religion* thus providing a much richer account of the way desires must be endorsed in order to be acted upon. We now turn to this.

2.2 The doctrine of *Gesinnung* and the hierarchy of maxims

According to Allison's analysis, the fully formed and explicated version of the Incorporation Thesis finally appears in the *Religion*. But its foundation is the first *Critique* notion of practical spontaneity outlined above and Allison is right to say that it is (at least) implicit in some of Kant's comments in the second *Critique*. In the fully developed Incorporation Thesis, agents are faced with *incentives* of two kinds: empirical incentives (incentives to do things which seem to offer happiness or omit things which seem to offer unhappiness) and duty. In order to act on an incentive, it must be accepted as an appropriate basis for action. In doing so, an

agent is said to *incorporate* that incentive into her maxim of action. Maxims are subjective practical principles or policies upon which actions are based. This act of incorporation is equivalent to the *adoption of a maxim*. The idea that an act of incorporation is required for action is clearly shown in a key quotation Allison (ibid., pp.39-40) provides from the *Religion*: there Kant says 'freedom of the will is of a wholly unique nature in that an incentive can determine the will to an action only insofar as the individual has incorporated it into his maxim (has made it into a general rule in accordance with which he will conduct himself . . .)' (R 6:24). As a rational power of choice it makes sense that a will requires a reason to incorporate an incentive into a maxim (i.e., to adopt a maxim of action). To repeat Korsgaard's example: I may take the strength of my desire as a reason for acting but as a rational being I must have a reason for doing so.

Kant sees the agent's choice of maxims as guided ultimately by his freely chosen fundamental maxim or 'meta-maxim'. He envisages two such overarching meta-maxims, sometimes called the moral maxim and the maxim of self-love. In the *Religion* (R 6:36), Kant argues that this supreme maxim must contain both the incentive of duty and 'the incentives of self-love and their inclinations' (R 6:36) because we are both moral beings and sensuously affected ones. As a result,

> the difference, whether the human being is good or evil must not lie in the difference between the incentives that he incorporates into his maxim (not the material of the maxim) but in their *subordination* (in the form of the maxim): which of the two he makes the condition of the other. (R 6:36)

In other words, it is not the case that the good have only the incentive of duty and the evil have only the incentives of self-concern in their meta-maxim. Rather, everyone has both of these in this maxim and within it, the good prioritize the incentive of duty over the concerns of self-love and the evil do the opposite. Korsgaard offers the following to give us an idea of what the meta-maxims[29] might be like:

> The maxim of self-love says something like:
>
> I will do what I desire, and what is morally required if it doesn't interfere with my self-love.
>
> and the moral maxim says something like:
>
> I will do what is morally required, and what I desire if it doesn't interfere with my duty.
>
> (Korsgaard,1996a, p.165)

---

[29] As the reader can see from the quotation, Korsgaard does not use the term 'meta-maxim' but instead uses the terms 'the maxim of self-love' and 'the moral maxim'.

The possession of a meta-maxim constitutes the agent's *Gesinnung* (usually translated as 'disposition'). *Gesinnung* is a central feature of Kant's conception of moral character since it guides the choice of maxims of action. We can already see in the second *Critique* the notion of a meta-maxim as the ground of the choice of lesser maxims of action in the following passage in which Kant is making just that claim with regard to the maxim of self-love:

> Now, a rational being's consciousness of the agreeableness of life uninterruptedly accompanying his whole existence is *happiness*, and the principle of making this the supreme determining ground of choice is the principle of self-love. Thus all material principles, which place the determining ground of choice in the pleasure or displeasure to be felt in the reality of some object, are wholly *of the same kind* insofar as they belong without exception to the principle of self-love or one's own happiness. (KpV 5:22)

Kant is more explicit about the idea that a freely chosen meta-maxim must be the ground of lesser maxims in the *Religion*. Having made the point that the subjective ground of the exercise of freedom must itself be freely chosen (or else no action can be imputed), he says,

> Hence the ground of evil cannot lie in any object *determining* the power of choice through inclination, not in any natural impulses, but only in a rule that the power of choice itself produces for the exercise of its freedom, i.e., in a maxim. (R 6:21)

This is closely followed by the point that,

> Whenever we therefore say, "The human being is by nature good," or, "He is by nature evil," this only means that he holds within himself a first ground (to us inscrutable) for the adoption of good or evil (unlawful) maxims (R 6:21)

According to Korsgaard (1996a, p.58), Allison (1990, pp.93-94) and throughout Caswell's article (2006a), this doctrine is not limited to an interaction merely between the meta-maxim which an individual adopts and his maxim of action. Instead, they say, Kant implies that 'between' these there can be a complex hierarchy in which each maxim justifies the adoption of a more specific one or ones below it, with maxims branching out in a downwards direction like the roots of a tree. So, for example, if I adopt the maxim of self-love as my meta-maxim, then this provides me with a reason to adopt a lower order (but still relatively general) maxim such as, 'I will only ever tell people things that will benefit me and nothing that will harm me', which in turn justifies still lower order maxims such as, 'I will lie when it suits me' and 'I will refrain from telling any truth which harms me'. It should be clear from this that in the practical philosophy, no non-moral act is the effect of some deep-seated natural impulse and that instead - even in cases of heteronomy - freedom goes 'all the way down' (or rather 'up') to a fundamental maxim and, we can say, beyond because the fundamental maxim is itself freely chosen by an 'act' of spontaneity. Since the meta-maxim is will's supreme maxim, it

has no higher maxim to guide it in its choice of that maxim. We shall examine how this fundamental maxim is chosen in the next chapter.

To complete the present exposition of the doctrine of the hierarchy of maxims: at the 'bottom' of the hierarchy are principles known as *Vorsätze*. These are situation-specific principles at which the understanding arrives from more general principles - maxims. This is done through another sort of merely intelligible activity of which an agent is in some sense conscious: practical judgement. It should be said that there is disagreement among commentators about what ought to count as a maxim, so some brief remarks on this are in order. Allison argues that there is a strong case for taking maxims to be principles which are more general than *Vorsätze* whilst not insisting (as he says Otfried Höffe, Rüdiger Bittner do) that all maxims as such must be sufficiently general to be 'life rules' or *Lebensregeln*, although the more general ones would be just this.[30] Allison makes the point that it may be difficult to distinguish between *Vorsätze* and the least general of maxims but the fact remains that there is a role to play for the understanding in judging how to apply an agent's more general principles to a situation.

3. The advantages of the Incorporation Thesis over Sidgwick's reading and Wood's

3.1 The Incorporation Thesis does not preclude evil

Since every action is based on a maxim which is chosen ultimately on the basis of the agent's *Gesinnung* and since this is also a *freely chosen* supreme principle, the doctrine of the hierarchy of maxims is a model of rational agency which goes hand in hand with the idea of transcendental freedom: instead of a limited conception of freedom in which we are only free to determine ends which will serve some unavoidable deep-seated natural impulse, Kant thinks we are, as it were, free 'all the way down' (from a practical rather than theoretical point of view) and this is true of an agent's actions whether moral or not.

The Incorporation Thesis, then, does more than *merely allow* us the notion of evil actions (by denying the Sidgwickian view that Kantian rational agency includes the notion of non-moral action determined by natural necessity). In its fully developed form as it appears in the *Religion* and *The Metaphysics of Morals* - as the doctrine of *Gesinnung* and the hierarchy of maxims - it also makes the agent's freely chosen moral disposition (whether good or evil) the

---

[30] Allison says these views appear in Höffe, O. (1979). Kant's kategorischen Imperativ als Kriterium des Sittlichen. In O. Höffe (Ed.). *Ethik und Politik* (pp.84-119). Frankfurt: Suhrkamp and in Bittner, R. (1974). Maximen. In G. Funke & J. Kopper (Eds.) *Akten des Kongresses* (pp. 485-498). Berlin: de Gruyter.

very basis of the model of deliberative rationality which he must freely *adopt* in order to act, since this disposition is itself a maxim. The practical standpoint of the first *Critique* is thickened (as Allison puts it) (1990, p.140) in these later works so that not only must the agent regard himself as free but his free choice must also be guided by a supreme *good* or *evil* principle in order to act at all. (This is what constitutes the so-called *Denkungsart* or 'way of thinking'.) Were Sidgwick's reading correct, its notion that heteronomous acts are caused by nature would have made evil impossible and rendered pointless the present project of illuminating moral self-development (since this consists in combating evil). However, since we can ascribe to Kant the Incorporation Thesis - a model of rational agency in which all acts are free and in which evil is not only possible but positively required[31] - we can rest assured that this project of understanding the nature of moral self-development is not rendered pointless by Sidgwick's misguided reading.

3.2 The Incorporation Thesis does not preclude moral development

There is a second interpretation of Kantian rational agency which rivals the Incorporation Thesis and which, if it were correct, would probably have made individual moral self-development impossible, thereby rendering a study which seeks to illuminate how this is to be done - such as the present one - also impossible. Again, fortunately the Incorporation Thesis rules this reading out as well. This is the view that the intelligible character of rational agency consists in reason acting as a sort of timeless, noumenal causality whose sensible effects appear in the phenomenal world as the actions of the human agent and that these actions constitute the empirical character of rational agency. Allison ascribes this interpretation to Allen Wood and claims that, in terms of the doctrine of *Gesinnung*, Wood's reading of Kantian agency would have us believe that, 'the causality of reason is to be conceived as operating by means of a single timeless choice of the intelligible character, which is itself the cause of the empirical character and its distinct, temporally appearing actions.' (ibid., p.50)

This picture is problematic for this study. Firstly, Kant believes that it is possible for an evil agent to bring about a moral 'revolution' in himself and this consists in rejecting the evil meta-maxim and choosing the moral one. Given that Wood's picture involves a *single* choice of intelligible character, it would seem to rule out such a conception of the revolution. Moreover, Wood (1984, p.94) considers 'drastic conversions from evil to good, or sudden degenerations from good to evil' (i.e., revolutions and, the opposite, *falls*) to be mere *phenomena*: events which one's life history might contain which can result from the choice of

---

[31] Assuming that not everyone takes the moral meta-maxim as the basis of the model of deliberative rationality they adopt for practical purposes: in short, assuming that not everyone is good.

a certain intelligible character. Obviously, this contradicts the view that such changes are themselves changes in intelligible character (i.e., involve choosing the other meta-maxim). In addition, Kant envisages the possibility of *inner* moral progress: for example, the acquisition of virtue which is the incremental development of strength of *will*. Such development would seem to be ruled out by Wood's picture. In fact, Wood does apparently regard incremental development as merely phenomenal. He says that 'temporal striving and moral progress' are 'the moving images of our eternal moral attitude, which we cannot conceive directly but to which we can relate only through such temporal images or *parables*' (ibid., p.99; emphasis added).

As Allison concedes, there is no doubt that much of the language Kant uses throughout his works suggests the idea of the intelligible character of rational agency as a timeless, noumenal *causality*. In one example which Allison provides from the first *Critique* Kant says,

> Reason is present in all the actions of men at all times and under all circumstances, and is always the same; but it is not itself in time, and does not fall into any new state in which it was not before. In respect to new states, it is *determining*, not *determinable*. (A556/B584)

He rightly points out that this (and other examples) suggest that the notion of timelessness is linked to the idea that reason operates as a direct causal power producing effects in the phenomenal world (Allison, 1990, p.48). We can find further examples of this which show that Allison's worry is well-founded. One such example is in a passage from the first *Critique* in which the issue of the compatibility of freedom and natural necessity is discussed at some length. This passage opens with the claim 'In respect of what happens we can only think of *causality* in two ways: either according to nature or from freedom.' (A532/B560; emphasis added) He goes on to argue that it is only through transcendental idealism that we can reconcile freedom and natural necessity. The argument is that if the phenomena of our agency were things in themselves there would be no freedom but

> If, on the other hand, appearances do not count for any more than they are in fact, namely, not for things in themselves but only for mere representations connected in accordance with empirical laws, then they themselves must have grounds that are not appearances. Such an intelligible cause, however, will not be determined in its causality by appearances, even though its effects appear and so can be determined through other appearances. Thus the intelligible cause, with its causality, is outside the series; its effects, on the contrary, are encountered in the series of empirical conditions. (A537/B565)

This quotation seems to support the position Allison ascribes to Wood with its reference to an 'intelligible *cause*' which apparently has effects in the temporal series. Allison's own examples

in a later chapter[32] (1990, pp.137-8) of this sort of worrying language include quotations from a key passage from the second *Critique* (KpV 5:97-100) in which Kant rehearses the theme of the first *Critique* passage which I have just quoted above - namely the relation between intelligible and empirical character. Here again Kant *might* be read as saying that one's entire phenomenal history is determined by a noumenal cause. He says,

> every action . . . even the whole sequence of his existence as a sensible being - is to be regarded in the consciousness of his intelligible existence as nothing but the consequence and never as the determining ground of his causality as a noumenon. (KpV 5:97-98)

He also provides a quotation from the *Religion* in which Kant says that the *Gesinnung* 'can only be one and applies universally to whole use of freedom.' (R 6:25)

Crucially, Allison claims that the Incorporation Thesis is a view of Kantian rational agency which whilst acknowledging Kant's frequent use of the term 'cause' in relation to free actions and the term 'causality' in relation to reason nevertheless allows Kant to avoid committing himself either to the view that reason is literally a causal power or to the related notion of timeless noumenal agency. He says that in Kant, to claim that reason is practical is to claim that reason provides rules for action which themselves have motivational force and because of this, 'It is certainly tempting, perhaps natural, to think of this force or determination in causal terms, or at least describe it in such terms.' (Allison, 1990, p.51) One might, he thinks, even go as far as thinking that Kant wishes us to understand principles of reason to be 'Humean causes of a particular sort - namely intelligible ones' (ibid.) which compete with empirical desires. However, we are not meant to take Kant's talk of the causality of reason to be an indication of a belief in the sort of literal causality described above. Instead, according to the Incorporation Thesis, incentives do not themselves cause action but are taken as acceptable bases of action by the agent. In addition, reason merely provides rules which influence our choices but do not necessitate them. All this means that,

> Although reason, according to this picture, is not literally an efficient cause of action, free actions are not regarded as uncaused. It is rather that the act of incorporation is conceived as the genuine causal factor and reason "has causality" only in the Pickwickian sense that it provides the guiding rule. (ibid., p.52)

According to the Incorporation Thesis, then, rational agents such as ourselves - that is to say, rational agents who have an empirical character in the sensible world - do not literally make a single choice in some timeless, noumenal realm, which choice causes effects across one's entire empirical existence. On the contrary, the claim is that in order to act, the agent - who

---

[32] Chapter 7 - '*Wille*, *Willkür*, and *Gesinnung*'.

has an empirical history and an empirical character - must adopt a model of deliberative rationality each time he makes a choice of an individual action, which model, by the time of the *Religion* at least, consisted in one's hierarchy of maxims headed by a meta-maxim and which constitutes one's *Denkungsart*. *This* is the way one's intelligible character pervades all of one's individual choices since the *Denkungsart* is brought to bear on each decision and thus has *influence* on them all (rather than simply causing them all as Wood suggests). The agent must adopt such a model because, as a rational being, he has no other way to make choices. One cannot simply sit and wait for nature to 'decide'. The only way for a rational being to act is for him to consult his policies and judge what action they recommend in the circumstances. This is just what it is to adopt a model of deliberative rationality or, if one prefers, to adopt the practical standpoint.

Furthermore, the adoption of this standpoint necessarily involves the exclusion of the third-personal or empirical one through which the world is seen in causal terms. In making use of his practical reason and consulting his hierarchy of maxims he frees himself from the causal principle of the Second Analogy. This is the sense in which the rational agent as such is *free from the condition of time of natural causation*. But this is not the same as having a literally timeless and noumenal causality making the corporeal agent's choices for him. When seen in this light, the claim from the first *Critique* quoted above that 'the intelligible cause with its causality is outside the series' (A537/B565) no longer seems the extravagant metaphysical claim it can be made out to be. It simply means that an act of incorporation requires a standpoint in which time and natural causality are not factors.

Similarly, in the last quotation from the second *Critique* given above, Kant seemed to be suggesting that an agent's entire phenomenal history is to be regarded as 'nothing but the consequence and never as the determining ground of his causality as a noumenon.' (KpV 5:97-98) However, in saying here firstly that the agent's history is a consequence Kant is merely pointing out that from the practical standpoint - i.e. 'in the consciousness of his intelligible existence' (KpV 5:98) - the agent's history consists in a series of determinations each and every one of which is a determination of a free will rather than the idea that there is a single noumenal cause with a set of many empirical outcomes. Thus, this point can be explained in a way which differs from the reading that says there is one noumenal choice with numerous, empirical results. The second point in the quotation - that the agent's history is never the determining ground of action - is the claim that regardless of the content of this history, every time he makes a choice of a maxim of action or an action based on it, it is as if that history does not exist. This is a point repeated in the *Religion* where he says,

> whatever his previous behaviour may have been, whatever the natural causes influencing him whether they are inside or outside them [*sic*], his action is yet free and not determined through

> any of these causes; hence the action can and must always be judged as an *original* exercise of his power of choice. (R 6:41)

The agent's history having, of itself, no power to determine the will is precisely the idea of the adoption of the practical standpoint in the manner of the Incorporation Thesis described above. It may seem strange to claim this history has no bearing upon his present choice but it only seems strange because new actions very often *do* follow historical patterns. But the point is when they do, it is because his negatively free, rational power of choice has continued to take his idiosyncratic inclinations as an acceptable basis for action. A habit's merely being a long-standing pattern of behaviour is *not in itself* sufficient for it to necessitate new actions in accordance with it since, for a rational being, no action is necessitated by anything. A new action which follows the pattern of an old habit is an old habit newly endorsed - albeit perhaps not explicitly or in full consciousness - and it is always possible to break it.

It is possible to reject the notion that an agent's *Gesinnung* is a noumenal cause rigidly having certain phenomenal effects whilst acknowledging that empirical character and the agent's outward behaviour will reflect somewhat his intelligible character. Firstly, we would expect an agent with a certain *Gesinnung* to tend to choose maxims of action according to recommendations of that meta-maxim, otherwise the notion of such a supreme maxim as an overarching guide would have no point. Secondly, if the intelligible character of the subject's rational agency is structured according to this hierarchy - by this set of maxims - we would expect to find certain patterns of behaviour in the phenomenal world which are a reflection of it and a reflection of the meta-maxim which shaped it. The thought, however, is that empirical character will only be 'roughly tuned', as it were, to intelligible character because actions do not follow as strict causal consequences of the agent's maxims in the manner of natural necessity. Although possession of a maxim is a necessary condition of the performance of the action-type to which it corresponds, mere possession of that maxim does not necessitate that action-type. In addition, there may be more (and in some cases perhaps many more) than one way of acting on a maxim. All of this is a far cry from the view that the agent makes a single inaccessible noumenal choice and that everything done in the phenomenal world follows ineluctably from this.

One important clarification regarding timelessness is in order. In addition to arguing that Kant does not regard human rational agency as somehow operating literally timelessly, Allison also quite rightly points out that, in any case, such a picture would be patently unsuitable for *sensuously affected* rational beings (which is what Kant takes human beings to be). In contrast, the operation of the power of choice in the manner envisaged by the Incorporation Thesis is suited to such beings. Presumably, he regards the latter as suitable because

although the power of choice is free from the conditions of time, as we have seen, a *sensuously affected* rational being must decide whether or not to act on *incentives* some of which (the empirical ones) consist only in feeling and this is obviously something experienced and in time. This means that, although not part of the causal order, this power of choice must be brought to bear on different choices at different times whenever they arise: for example, I may have maxims concerned with the maintenance of my personal computer now but I did not and could not have had them before personal computers were available to the general public. Similarly, the whole notion of learning new skills relies on the possibility of adopting *new* maxims which I did not possess *before*. This does not detract in any way from the idea that the power of choice which adopted them did so free from the condition of time of natural necessity but instead in a merely intelligible 'moment' of spontaneity. It is in this sense that the choice of a particular maxim of action is 'timeless'.

4. The adoption of meta-maxims and 'timelessness'

However, in contrast to this, Allison denies that the choice of a *meta-maxim* is something which happens at a specific time. For example, when explaining what *Gesinnung* is he says 'by stressing the analogy between *Gesinnung* and ordinary maxims, we have ignored the significant difference between them . . . although acquired, a *Gesinnung* is not acquired in time.' (Allison, 1990, p.143) Later he elaborates on how we should understand this 'timeless' acquisition (of the evil meta-maxim) saying,

> it is not to be thought of as explicitly and self-consciously adopted by an agent. . . . *It is rather that one finds that this is how one has been behaving all along.* Nevertheless, since this behaviour is that of a rational agent . . . it must be thought of as involving action based on the "conception of law" and, therefore, a maxim. (ibid., p.153; emphasis added)

In his article 'Perversity of the Heart', David Sussman expresses an interpretation of the adoption of a meta-maxim (which he ascribes[33] to Korsgaard in 'Morality as Freedom'[34]) which is remarkably close to Allison's since like the latter he rejects both the idea that it 'takes place in some sort of timeless noumenal realm' and that it is 'an act that can happen at any particular point in time'. (Sussman, 2005, p.172) Sussman thinks that the conduct of one's entire life amounts to the adoption of a meta-maxim and he compares this with the way one might go for a walk without having made a prior conscious decision to do so. He says,

---

[33] Correctly, I believe.

[34] Reprinted in her *Creating the Kingdom of Ends* (1996a).

> Instead, this intention is to be found in the way I walk, something that is not ultimately distinct from the activity itself. In coming to walk, I simultaneously come to have the intention to walk: each is an integral aspect of the same overall activity.' (ibid.)

To some extent, it is fortunate for the present project that this interpretation of the adoption of a meta-maxim is available since Kant takes the revolution from an evil disposition to a good one as a necessary condition of moral progress and, as mentioned earlier, the view which takes the adoption of a meta-maxim to be literally timeless would make it very difficult to see how such a revolution could take place.[35] Sussman's interpretation therefore removes one obstacle to this important idea. However, it creates another one by insisting that the conduct of one's *whole life* adds up to a single meta-maxim. Presumably, the idea of the revolution requires that more than one is possible since it seems implausible that (for example) in successfully overturning evil in the latter part of my life I make the earlier part of my life one in which I was guided by the moral meta-maxim.

Sussman says that it is 'the moral biography' or 'the course of a life' (ibid.) or 'the shape of my moral life as a whole' (ibid., p.173) that counts as the adoption of a meta-maxim. He also says that (in the case of the moral meta-maxim at least) what amounts to the choice is 'a *commitment* to morality' (ibid.; emphasis added). I agree that the adoption of the moral meta-maxim is seen as a commitment in the key passage[36] from the *Religion* on the revolution upon which both Sussman and Korsgaard are relying[37] (R 6:48). However, if it is a commitment which counts as the choice of a meta-maxim, then assuming that commitments can *change*, presumably I can have more than one meta-maxim in my lifetime - by being committed to the willing of evil maxims of action across one part of my life and to good ones across another. To modify Sussman's analogy: just as in coming to walk, I form the intention to walk, later, in coming to run, I form the intention to run. That my walk turned into a run does not make it the case that I was always running, nor does my later intention to run somehow make it the case that my earlier intention was also to run rather than walk. It is still the case now that my intention *then* was to walk.

5. Conclusion

We have seen that Kant's true view of rational agency - enshrined in the Incorporation Thesis - is such that evil is possible since, according to that view, heteronomous acts are negatively

---

[35] At least from the agent's point of view as opposed to God's and it is the former which is important for the purposes of this study.

[36] We will examine this passage in Chapter 6 when we discuss the revolution.

[37] Korsgaard quotes the entire passage on the revolution (R 6:47-48) and Sussman explicitly states that his views on this follow hers (and in my judgement, do so).

free just as autonomous ones are and so agents are responsible for failing to act morally. The present project of illuminating moral self-development is not precluded or rendered pointless by any supposed impossibility of evil within the Kantian framework. We have also seen that it is not the case that Kant supposes that the intelligible character of rational agency should be characterized as a single, timeless noumenal choice which ineluctably 'causes' the choices of the human being. Instead, it is seen as a model of deliberative rationality, which the agent adopts every time she must make a choice. Given this, it is therefore also not the case that 'moral progress' could only ever be the mere outward unfolding of the results of the fundamental choice, as Wood thinks. Instead, there is nothing about Kant's conception of rational agency which would prevent the agent from self-consciously making particular choices to engage in true, *inner* moral development (for example, in the form of the acquisition of virtue or the purification of his motives.) In addition to the *operation* of rational agency not being literally timeless, the 'timeless' *acquisition* of the meta-maxim which forms the basis of this agency can be understood in a way which is friendly to a study of moral development in Kant. Finally, since moral development consists in the combating of evil, the interpretation of Kantian evil we adopt is important. For this reason, it is again fortunate that the Incorporation Thesis is the correct interpretation of Kantian rational agency since its basic elements - the presentation to the will of incentives and the adoption of maxims - are foundational to the particular interpretation of evil I wish to endorse in the next chapter.

Chapter 3

Radical evil

In order to understand how it is possible to develop as a moral agent within the framework of Kant's practical philosophy - in order to see what agents must actually do - I take it as necessary first to understand the nature of the problem which such development is meant to address, that is to say, the nature of Kantian radical evil. This is no straightforward task as the account given in the *Religion Within the Boundaries of Mere Reason* is problematic in a number of ways which we shall explore presently. These difficulties are raised by Seiriol Morgan in his article 'The Missing Formal Proof of Humanity's Radical Evil in Kant's *Religion*', and are addressed by his rational reconstruction of Kantian radical evil. As we shall see, another analysis plays down some of the difficulties which the *Religion* presents but I believe that the problems are real and Morgan's solution is required. In addition, I consider David Sussman's attempt to provide an 'anthropological' account of the emergence of evil through a natural process. We will see, I think, that this attempt fails. Hence, I endorse Morgan's reconstruction. When we come to consider how a person may develop as a moral agent later in this study (Chapter 6), we will have arrived at a version of Kant's account which, unlike the original, does not take all of mankind to be universally evil. Instead, it sees everyone as either having endorsed evil or as tempted by it. In addition, the good are seen as committed to morality but as capable of failing to live up to this commitment or even abandoning it altogether. The view of the human propensity to evil involved in this picture does not appear in the corpus but is constructed from materials from it, and vindicates many of Kant's claims concerning it.

1. The predisposition to good, evil and good dispositions and the propensity to evil

Though the reader may be familiar with it, it may be useful to begin with a brief summary of Kant's own account of evil from book I of the *Religion*. To begin, Kant believes (R 6:27-30) we have a 'predisposition' (*Anlage*) to good which has three aspects: Firstly, *animality*, which consists in the instinctive inclinations to acquire the basics of survival, reproduction and society; secondly, *humanity*, in which we use our practical reason to pursue our desires and to assess our level of happiness; thirdly, *personality*, in which the moral law is seen as an incentive sufficient for the will to adopt moral maxims. Kant thinks these predispositions are essential to human beings; it is impossible to be a human being without them (R 6:28).

In addition, as we saw in Chapter 2, there are two dispositions (*Gesinnungen*)[38] both of which are at least available to an agent: if the agent prioritizes the incentive of duty over that of self-love, she thereby adopts the moral meta-maxim and has a good disposition; if she prioritizes the incentive of self-love over duty, she thereby adopts the maxim of self-love and has an evil disposition. As we saw in the last chapter, our choices of maxims of action are grounded in higher order maxims in the hierarchy and ultimately in the meta-maxim. Kant is a disposition-rigorist: everyone must either have a good or evil disposition but it is impossible to have both and impossible to have neither. However, Kant's view is that although the acquisition of a good disposition is possible through a revolutionary act of will, every human being has made a free choice to be evil.

Finally, we have a propensity (*Hang*) to evil. Kant regards this as universal: although it cannot be inferred from the concept of a human being, he thinks that because of the many examples of evil in the world, 'we may presuppose evil as subjectively necessary' (R 6:32). However, he also says that in view of these 'woeful examples', 'We can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being' (R 6:32-33), which suggests that such a proof is available. Kant thinks the propensity to evil is imputable to us because he takes it to be rooted in our freedom. Since he also thinks it is 'the formal ground of every deed contrary to the law' (R 6:31) it seems to be doing the work we would expect the evil meta-maxim to be doing in the theory. It may seem, then that Kant identifies the propensity to evil and the evil disposition understood as the evil meta-maxim. Henry Allison certainly thinks they are the same on the grounds that they both determine 'the orientation of one's *Willkür* as a moral being.' (Allison, 1990, p.153) And whilst Matthew Caswell says that 'it is not necessary to claim that the notions of *Hang* and *Gesinnung* are identical for Kant,' he also says that 'these descriptive aspects belong to a single element within Kant's theory of finite rational agency.' (Caswell, 2006a, p.199) At any rate, we can see that the orthodox Kantian view is that the universal propensity to evil is at least based on or is an expression of the universally adopted maxim of self-love.

According to Allison (1990, p.155) Kant's reasons for thinking we all have a propensity to evil rather than good is that if anyone had the latter, he would be a finite rational agent for whom the choice of moral action would never be a struggle. Having a propensity to good would be the result of having a good disposition. Such an agent - one who did in fact prioritize duty over self-love - would incorporate duty into *all* of their maxims of action: the thought here is that there are no half-measures for the one who has genuinely accepted the majesty of the law. However, since moral action is - at least from time to time - arduous for everyone, it cannot be the case that anyone has a propensity to good and this means no one has a good

[38] The singular form is *Gesinnung*.

disposition, so, given rigorism, we must all have evil dispositions and therefore a propensity to evil. So, the intuitive notion that everyone is susceptible to the propensity to do evil - that evil is possible for everyone - gives rise to the decidedly unintuitive notion that everyone has an evil disposition.

There are three levels to the propensity to evil: First, frailty, in which duty is a weaker incentive than the inclinations that present themselves even though the agent has adopted a moral maxim; second, impurity, in which the agent finds insufficient motivation in the incentive of duty to perform the act prescribed by the moral maxim he has adopted and he must rely on supplementary inclinational motivation to act on the maxim; and third, depravity in which the agent prioritizes the incentive of self-love over the incentive of duty in all his choices. The frail and impure are said to have good dispositions, the depraved, an evil one.

2. Morgan and the propensity to evil as incentive to license

As we have just seen, Kant infamously says (R 6:32) he may spare himself the formal proof of our 'corrupt propensity' (which he seems to regard as available), instead taking the multitude of examples of evil deeds found in experience as the right kind of evidence for condemning mankind as (subjectively) necessarily evil. As Morgan points out (Morgan, 2005, p.68), crucially, this is not merely a claim that we all have a susceptibility to do evil things; the claim is that we are all radically evil, as we saw above. This type of evidence, which Kant provides, is of course insufficient to establish such a claim. As Morgan notes, all it can do is establish that there are evil individuals or perhaps that evil is commonplace (ibid., p.65). Kant is not entitled to conclude from it that there is a universal propensity to evil and that we are all evil people. What is needed, he says, is a transcendental deduction and his rational reconstruction is an attempt to provide this. As we shall see, it also provides a way to vindicate many of the claims Kant wishes to make for the propensity to evil: that it is universal, imputable and inextirpable. This it does by locating such a propensity to evil in the very nature of a free will as pure spontaneity.

2.1 The will can have reasons *qua* free will - the argument from spontaneity

Morgan seeks to reconstruct Kant's account of evil out of materials available in works prior to the *Religion* and finds his starting point in an idea from the *Groundwork*, the importance of which will be explained when we move on to Morgan's theory proper. This starting point is the idea that 'the will can have reasons simply qua free will and that any such reason is a function of its own nature.' (ibid., p.79) This is shown in Morgan's representation of the argument of *Groundwork* III, which now follows. Kant thinks that, as a kind of 'causality', a

free will must be subject to a law or is otherwise an absurdity (G 4:446): its 'choices' would be random and hence not reasoned choices at all. But since the will is free, this law must be one of freedom and he takes the categorical imperative to be this law. The reasons why it must be this law are found in Kant's appeal to our membership of the intelligible world in *Groundwork* III. In considering the difference between our passivity or receptivity in sensation versus 'our pure self-activity' (G 4:452) in the production of ideas, we realize we can regard ourselves from two standpoints (that of the phenomenal world in which we regard ourselves as passive parts of the causal order and its law and that of the intelligible world which is governed by laws of reason). A rational being '*qua* intelligence' must regard itself as a member of the intelligible world. We must then regard ourselves as pure spontaneity, 'since anything with respect to which we are not spontaneous but receptive comes not from the free power of reason but the passive effect of sensibility.' (Morgan, 2005, p.75) But only autonomy - being a law to ourselves - allows the expression of our freedom because the only other alternative is a principle through which we would be determined in the manner of natural necessity. This means the principle of a spontaneous will must be a principle of autonomy. But as Morgan says, 'the only principle of autonomy is the formal one of the Categorical Imperative' (ibid.). So the spontaneity of reason commits us to morality. Although we can conceive of ourselves as members of the phenomenal world, Morgan says Kant thinks we must regard ourselves 'as belonging more fundamentally to the intelligible world' since 'the intelligible world contains the ground of the sensible world' (ibid.). According to the first *Critique*, causality is not a feature of the world itself but instead is one of the pure concepts of the understanding through which the mind synthesizes the manifold of experience, making sense experience possible. Morgan says, 'His thought seems to be that our experience of the world of causality, and the sensible inclinations it throws up in us, depends upon the spontaneous activity of the mind as prior to it.' (ibid.) And so 'the intelligible, purely spontaneous aspect of the self' (ibid.) is thought to be the 'proper' self. The pursuit of inclination is an abandonment of this self and its essence, freedom. So, the preservation of freedom through the adoption of the principle of morality, i.e., that of autonomy is one 'trumping any consideration that sensibility might, on the face of it, seem to present us as a reason for choice.' (ibid., p.76) Kant takes this argument to have shown the 'universal rational authority of morality.' (ibid.)

Since he thinks Kant's *Groundwork* III argument might not be very clear, Morgan presents a thought experiment (ibid., pp.76-78) (borrowed from Christine Korsgaard)[39] to explain why reason as pure spontaneity has overriding reason to take morality as its supreme practical

[39] This explanatory conceit is from Korsgaard (1996b, pp.97-98, pp.219-33). She takes the choice of a meta-maxim to be something which does not take place straightforwardly in time. This is a difficult notion and the analogical use of temporal language which the conceit as such licenses clarifies that notion: it is *as though* the will chooses a meta-maxim 'before it enters the world'.

principle and that this reason is located in its own spontaneity and so it can have reasons *qua* free will. In this conceit, we imagine a pure, rational will trying to choose its supreme reason-giving principle 'before it enters the world'. Since it is a will and hence, free, the imposition of this supreme principle by nature is excluded. Morgan says if, for the sake of argument, we limit ourselves to Kant's strictures, there would only be two choices for a supreme reason-giving principle: self-love or morality. The will must therefore decide whether it will prioritize the satisfaction of inclinations or conform to the categorical imperative. This amounts to the choice of the good or evil disposition.[40] No appeal to morality or self-love is available without circular justification. This will has no principle to guide it in its choice because the principle it is choosing here is supreme. One might think then that it would simply plump for a principle randomly. But if this supreme principle is to be reason-giving for the choice of all future principles, which are themselves supposed to be reason-giving, then it and therefore they would not be reason-giving if it were 'chosen' in a volitional spasm. This will, with nothing else to guide its choice, only has its 'sheer power of choice, the will's spontaneity' (ibid., p.77) for this purpose. Since the will is freedom itself, 'the only thing that can possibly provide the will with a reason is spontaneity itself.' (ibid.) A choice of self-love would limit its willing to following happiness through the satisfaction of inclinations. Since these are given to us by nature and since their satisfaction consists in acting as though we are unfree, (i.e. as though we are part of the causal mechanism of nature), such a choice would be an 'abrogation of freedom' (ibid., p.78). One might think that the constraint that the categorical imperative demands does not seem like freedom. However, this can be understood as freedom's constraint of itself; hence this constraint is what freedom itself consists in. I would say, adherence to this law is the freest a will can be whilst still having a reason-giving principle, the possession of which prevents it from being an absurdity. The will, then, has overriding reason (i.e., incentive) to choose morality.

2.2 The incentive to license

However, Morgan says that '*we know that at least some wills do not choose morality*' (ibid., p.79). Presumably this is because we know, for example, from their repeated and manifestly abhorrent actions that, within a Kantian framework, they could only have the evil meta-maxim as their supreme guiding principle (and one can and must have either the moral meta-maxim only or the evil one only according to Kant's rigorism.) It is hard to understand such a choice given the arguments presented in the previous sub-section. As Morgan puts it, the will can have no reason to adopt evil (since any 'reason' which is opposed by an overriding one is no reason at all) (ibid.). He also points out that appealing to the draw of inclinations themselves does not help to understand the possibility of wrong-doing because what we are

[40] See Section 1 of this Chapter.

looking for is what it is about the will that gives it a 'reason' to adopt an overarching principle through which it sees these evil desires as being choice-worthy (ibid.).

Morgan thinks the will itself must be the source of this pseudo-reason. As we saw above, the spontaneous will can take the expression of its own spontaneity as overriding reason to adopt morality as its supreme principle. And so, Morgan takes the argument of the *Groundwork* to have established (among other things) that 'Kant thinks that the will can have reasons simply qua free will and that any such reason is a function of its own nature.' (ibid.) So, this opens up the possibility that there is some pseudo-reason-giving feature of the will itself, through which it recommends to itself the adoption of a principle of preferring self-love. He says, 'the pseudo-reason must lie in the will's erroneous representation of freedom itself, in such a way that it is tempted in pursuit of this freedom to make a choice that in fact self-destructively renounces it.' (ibid., pp.79-80)

Thus, the question of what this erroneous representation of freedom is arises. Morgan believes a clue to this can be found in Kant's distinction between negative and positive freedom. Morgan defines negative freedom as 'simply the power of activity in the absence of alien determination'[41] and positive freedom as 'autonomy in the form of the Categorical Imperative.' (ibid., p.80) A will which takes its negative and positive freedom to be as defined above would affirm its spontaneity by adherence to the laws of freedom - the moral law. However, a will which fails to be conscious of its positive freedom would have an incomplete conception of freedom. Morgan wonders how a will which took its negative freedom to be the whole of its freedom would affirm its spontaneity. If negative freedom, in this case, were construed correctly - as freedom from alien determination - this would not provide the will with a reason to affirm its spontaneity through self-love. What would provide this reason is if the will were to find its 'negative freedom presenting itself not in the form of lack of alien determination, but *lack of restraint*' (ibid.). Correlatively, if a will accepts such a conception of negative freedom, it would take the affirmation of its spontaneity to consist in doing whatever it wills to do. This representation of freedom as lack of restraint and as simply doing what it wills to do, acts as an incentive to adopt just that principle 'simply to do what it wills to do' (ibid.). The following considerations should explain why such a principle is identical to the meta-maxim of self-love.

---

[41] Kant's definition of negative freedom in the *Groundwork* is: '*freedom* would be that property of such a causality [the will] that it can be efficient independently of alien causes determining it . . . . The preceding definition is *negative*' (G 4:446).

We all pursue what Kant calls outer freedom (freedom from external restrictions on action in the phenomenal world). One important reason for this is that doing things in the phenomenal world is necessary in order to attend to our needs and as Kant says in the second *Critique*,

> The human being is a being with needs, insofar as he belongs to the sensible world, and to this extent his reason certainly has a commission from the side of his sensibility which it cannot refuse, to attend to his interest (KpV 5:61)

Pursuing outer freedom is not in itself wrong unless such pursuit involves trampling on the rights of others. However, Morgan says this is just what the will which has endorsed the incentive to license does. To a will which fails to acknowledge that its freedom lies in the self-constraint of autonomy and which succeeds in taking its freedom as consisting in lack of all restraint and in doing as it wills, it seems that it can only affirm its spontaneity by pursuing ends of self-interest, pleasure and the avoidance of harm and pain and this entails being as free as possible from the self-constraint of the law and from external constraints in the world of sense. Such a will takes the untrammelled pursuit of outer freedom to be freedom *simpliciter*. In fetishizing the outer freedom of the world of sense as legislative over the inner freedom of autonomy, its policy is identical to the maxim of self-love. Morgan calls the misrepresented freedom which provides the will with a reason to adopt this principle, *the incentive to license*.

2.3 Vindication of Kant's claims for the universality, imputability and inextirpability of the propensity to evil

Morgan's reconstruction preserves and vindicates the features of *universality*, *imputability* and *inextirpability* of Kant's propensity to evil. First, *universality*: this is a transcendental deduction of radical evil (Morgan, 2005, p.86) because it explains the conditions of the possibility of incidents of wrong-doing to be found in the world of sense. Certain evil acts indicate possession of evil maxims of action and an evil meta-maxim. The choice of such a meta-maxim *seems* impossible given the will's overriding reason to choose morality *unless* it had an incentive to do so. But at the 'top' of the hierarchy the only thing that can provide a reason is whatever the will takes to allow the expression of its spontaneity. So any incentive must be one which promises (albeit falsely) to allow this expression. The only thing that could do this is the will's misrepresentation of its own freedom as consisting in lack of all restraint[42] and in the unlimited pursuit of outer freedom, (the incentive to license). There must be such an incentive, or else the free endorsement of evil would be impossible. Since the incentive arises from the will's own nature as spontaneity which is a feature shared by all wills, the incentive

[42] By the law and by other wills.

is present in all wills. Thus Morgan has vindicated Kant's claim that the propensity to evil is universal.

Second, *imputability*: in the *Religion*, Kant wants to claim that having a propensity to evil is itself imputable to us which means he must think of it as freely brought upon ourselves. But as we have just seen, he also thinks of it as universal. The problem here is that if something is universal, then it might be hard to see how it could be regarded as freely chosen and indeed, this is an issue Kant neglects in the *Religion*. However, as Morgan says, 'the problem can be resolved if we take the propensity to evil to lie in the self-assertive tendency of the will' and that 'the temptation is something that the will offers to itself.' (ibid., pp.91-92). Since it emerges from the free will itself, it cannot be the result of natural causation as this would contradict the will's own essence as freedom. Having a propensity to evil is simply something free wills as such inevitably and freely choose to have. Morgan points out that some might object that if the propensity is universal, it is unavoidable and hence, not imputable. However, this assumes that imputability can only be had through the 'liberty of indifference', which conception of freedom is 'the power to do otherwise than one actually does' (ibid., p.92). Morgan argues that Kant rejects the liberty of indifference as either necessary for or constitutive of freedom. As we have seen, in Morgan's account of the argument in *Groundwork* III and in Korsgaard's conceit, the essence of the will is its spontaneity. So, the freedom of the will consists in its 'liberty of spontaneity', that is to say, 'to be oneself the cause of one's own actions' (ibid.). Since the incentive is in this way an originary upsurge from freedom itself, it is imputable to all wills.

However, one might wonder how a will which fails to conceive of its freedom as autonomy could be considered subject to the moral law since it is in autonomy that the laws of freedom consist. I think the answer is that the law still applies to this agent because (by Kant's lights) he is *in fact* autonomous (in the sense of possessing a will which has the property of autonomy) but he is wilfully refusing to be conscious of this (at least at the level of the meta-maxim). Since the licentious will sees the affirmation of its spontaneity in license and since a return to consciousness of a full conception of freedom can curtail this, such a will takes an interest in preserving this state of affairs even though the denial of the exercise of positive freedom is in fact inimical to the true spontaneity of the will. However the will can be conscious of its positive freedom, so we are permitted to say it ought to. We will examine this in more detail in the next chapter when we consider the involvement of self-deception in facilitating unfreedom in a free will. Third, the account vindicates *inextirpability*: since the evil 'incentive emerges from its [the will's] innermost nature in this way, it does not seem unreasonable to consider it inextirpable'. (ibid., p.87)

2.4 Resolution of an inconsistency in Kant's account of good and evil dispositions and the propensity to evil

As we noted earlier, Kant seems to take the propensity to evil and the evil disposition as both consisting in having the evil meta-maxim.[43] With Morgan's formal proof we now must see them as distinct aspects of the will since the propensity is now thought to consist in *the incentive to adopt* that meta-maxim - understood as a policy of licentiously pursuing outer freedom - but not to consist in *possessing* that meta-maxim itself. It is fortunate that the formal proof separates the notions of propensity to evil and evil disposition since as we shall see now, regarding them as the same gives rise to an inconsistency in Kant's account of evil.

Morgan argues that the inconsistency occurs in the following way: as noted above, Kant claims that the propensity to evil is universal, consists in the adoption of an evil disposition and is inextirpable. But Kant also claims that agents with an evil disposition can undergo a 'revolution' and adopt a good disposition. Morgan says, 'This would imply that some human beings can come to possess both good and bad dispositions at the same time.' (ibid., p.94) The problem for Kant, he points out, is that this contradicts disposition-rigorism, since that doctrine says one must either have a good or evil disposition but cannot have both (or neither). This means we must drop one of the claims above.

If Morgan is to stand by his own rational reconstruction, he cannot abandon either the claim to the universal propensity to evil or to that of the inextirpability of it. Also, he rightly says that rejecting the idea of the possibility of a revolution in disposition 'seems a recipe for despair and threatens to annihilate moral responsibility.' (ibid., p.95). Rejecting rigorism would require endorsing one of the two forms of 'latitudinarianism': indifferentism (the view that the human being is by nature committed to neither good nor evil) or syncretism (the view that she is both at once). Morgan points to Kant's rejection of both of these in the *Religion* (R 6:22-25): in claiming that a will may be neither committed to good nor to evil, indifferentism thereby denies the will a way to incorporate incentives according to principles it has adopted. This goes against Kant's entire conception of rational agency comprising the Incorporation Thesis and the hierarchy of maxims as laid out in Chapter 2. Endorsing indifferentism would modify the practical philosophy beyond all recognition.

Morgan (2005, p.96; p.112, n.18) explains the problem with endorsing syncretism by drawing our attention to the thought experiment (explained in Sub-section 2.1). Recall that the spontaneous will is 'trying' 'before it enters the world' to choose a fundamental principle with which to guide its further choices. The will which takes its full freedom to be autonomy will

---

[43] Morgan says, '(at R 6:43) Kant explicitly states that the propensity to evil involves the adoption of an evil supreme maxim.' (2005, p.68)

see (rightly) that only the moral meta-maxim[44] allows it true expression of its spontaneity and recognizes that it therefore has overriding reason to endorse this maxim as its supreme practical principle. The evil will, on the other hand, misconstrues its freedom as freedom from all constraint and as the licentious pursuit of outer freedom and so takes it that it has overriding reason to adopt the maxim of self-love: a policy of not constraining itself by the law nor willingly submitting to the restraint of others. In both of these wills (good and evil), any compromise with the opposite policy would be an *entirely unjustified abandonment* of a policy it has accepted because of what it takes to be an overriding reason: namely that the policy it has chosen is the only one which allows the expression of its spontaneity as it sees it. Since rigorism is the only other available view of disposition, it must be retained.

This means that the inconsistency outlined above can only be resolved by seeing the propensity to evil and the evil disposition as separate aspects of the will: this separation allows us both to see the propensity as universal and inextirpable *and* without inconsistency we can allow that an agent can come to possess a good disposition without his retaining an evil one - i.e., without violating rigorism. In separating the propensity to evil and the evil disposition, the inextirpability of the former does not entail the inextirpability of the latter.

It is important to note that endorsing disposition-rigorism, however, does not commit Morgan to the notion of maxim-rigorism - the view that, for example, a will with a good disposition (i.e. meta-maxim) must as such always endorse good maxims of action. Having a good meta-maxim is a commitment to an ideal of morality. But it is entirely consistent with such a commitment that there be some failures to live up to one's ideals. In addition, it seems maxim-rigorism would mean that an evil will could not endorse the good maxims, which we will see in Chapter 6 are required for the revolution from evil to good, a revolution which Kant takes to be possible (because it is required). For this reason he says, 'it certainly cannot be self-love, which when adopted as the principle of all our maxims, is precisely the source of all evil.' (R 6:45) The way such exceptions seem to work is that the good will, for example, takes autonomy as its conception of freedom broadly speaking but in relation to certain individual desires may incorporate them and the licentious conception of freedom into individual

---

[44] Korsgaard says that the moral meta-maxim (which she calls the moral maxim) 'is the maxim derived from the Categorical Imperative' (1996a, p.165). All wills, in a sense, 'choose' the categorical imperative because all wills are in fact autonomous. Perhaps a better way of putting it is that the categorical imperative just is the supreme practical principle of any will in virtue of the fact that it is autonomous. But this is consistent with an individual will's choosing the *evil* meta-maxim as its supreme *subjective* practical principle, since, although it actually has autonomy (and is subject to the categorical imperative), it has embraced the temptation not to conceive of its freedom *as* autonomy and hence falsely supposes that its spontaneity is best expressed through lack of any constraint (either its own or anyone else's) and so can (despite actually having overriding reason subjectively to choose autonomy) pursue outer freedom licentiously.

maxims. We will discuss the notion of exceptional maxims further in the sequel where self-deception seems to play a role in their possibility.

2.5 The intuitive appeal of the incentive to license

Perhaps the most valuable feature of Morgan's reconstruction is that it provides a formal proof of an aspect of the will which Kant regards as universal and 'subjectively necessary' (R 6:32). As we have just seen, it also eliminates an inconsistency in Kant's argument by separating the propensity to evil and the evil disposition. In addition, as Morgan points out, through such separation, it addresses Kant's unintuitive conflation of a propensity - which, arguably, we would normally understand as a '*susceptibility* to a kind of behaviour' (2005, p.97) - and a disposition - which we perhaps would normally understand as a *state*. As Morgan says, 'someone who frequently resorts to violence when angered might appropriately be described as having a propensity to violence, but someone who is married is just married.' (ibid.) To extend Morgan's example, we might meaningfully say someone has a propensity to marry when proposed to - i.e., is constantly faced with a temptation to marry and sometimes or often gives in to that temptation thereby entering into the state of marriage on those occasions.[45] But the state of being married is not *in itself* this propensity (even if it might be an indication of it).

In terms of Kant's conceptual framework, I believe the way Morgan's reconstruction allows us to think of the propensity to evil as a susceptibility is by characterizing that propensity as in *incentive* to license: in Kant, an incentive is a temptation which presents itself to the will but, as we saw in Chapter 2, there is nothing which necessitates the endorsement of an incentive - it may only determine the will if freely incorporated into a maxim. Similarly, the incentive to license must be endorsed into one's supreme maxim to be effective, since only then do we adopt an overarching evil principle, which guides us to endorse lower order empirical incentives and thereby adopt maxims of action which serve evil. When it presents itself, we may decline it but as long as it is present, it is 'waiting', as it were, to be embraced.

Morgan takes Kant to have identified the propensity to evil and the evil disposition because, for Kant, they both consist in the adoption of the evil meta-maxim. Kant's motivation to do this, he argues, is that he takes it as a theoretical constraint that both must be imputable and this involves seeing the propensity (as well as the disposition) as a freely chosen meta-maxim. However, there may be a way to see this as not constituting grounds for an identity claim. In his article, 'Kant's Conception of the Highest Good, the *Gesinnung*, and the Theory of Radical Evil', Matthew Caswell presents an analysis, which attempts to show how the

[45] Assuming the proposals are sincere and so on.

propensity to evil and the evil disposition can both be considered to consist in the adoption of the evil meta-maxim but that the two have distinct roles in the theory. His analysis is also an attempt to preserve the idea of the propensity as a tendency to evil and to elucidate Kant's unintuitive notion that we are all evil. If successful, this analysis would demotivate the perceived intuitive need to separate the propensity to evil and the evil disposition, as Morgan's reconstruction does, (although it would not remove the need for a formal proof, which Morgan has provided). I shall argue that his analysis either involves an unintuitive conception of the possession of the good meta-maxim as sufficient for moral perfection or must separate the propensity to evil from the evil disposition as Morgan's does.

Before we examine Caswell's account of the propensity to evil and the evil disposition, it is necessary to explain the conception of possession of the moral meta-maxim he is working with. He asserts that Kant believes 'a good *Gesinnung*,[46] as a fundamental commitment[47] to direct one's life according to the moral law, bars the adoption of any particular evil maxims - a sound tree cannot bring forth evil fruit.' (Caswell, 2006a, p.198) This is the notion of maxim-rigorism - the view that all particular maxims of action 'line up' with the meta-maxim one has endorsed. On this view, a good *Gesinnung* consists in moral perfection. He then says that Kant makes 'the claim that evil is universally attributable' (ibid.) and says this may be seen as harshly counterintuitive. However, he thinks he can provide an analysis which 'will show that his view is not so hideous' (ibid.).

Let us first see if Caswell's account can demotivate Morgan's notion that the propensity to evil and the evil disposition ought to be entirely distinct. Caswell takes the notions of a moral[48] propensity and a moral disposition as 'aspects' of the possession of a meta-maxim. Perhaps Morgan's separation of the propensity to evil and disposition could be demotivated if we can see *how*, according to Caswell, they are both aspects of one thing: having an evil meta-maxim.

He gives an account of what he means by the propensity to evil as an aspect of the possession of an evil meta-maxim. In contrast to a person who has adopted the moral meta-maxim (who is morally perfect) those who have the evil meta-maxim, have a propensity to evil, understood as a *susceptibility* to evil, systematically built in, as it were, into their entire

---

[46] At this stage of his article, Caswell is taking '*Gesinnung*' to mean 'possession of a meta-maxim'.

[47] Caswell's linguistic intuition concerning the meaning of 'commitment' clearly varies from mine since, (as I explained earlier) I believe that it is precisely because the possession of a meta-maxim is a commitment (to good or evil) that it admits of the willing of countervailing particular maxims.

[48] It is clear from the context (Caswell, 2006a, p.198) that by 'moral propensity' and 'moral disposition' he means respectively 'propensity to do with morality' and 'disposition to do with morality' rather than 'propensity to good' and 'good disposition'.

deliberative bent. Another way to put this is that, for these people, moral fallibility is at the very root of their practical deliberation: evil is *possible* for them, unlike those with a moral meta-maxim (on Caswell's view at this stage). People who we may think of as generally good 'have their limits'. To further illustrate this, Caswell refers to Kant's quoting the British parliamentarian who said ' "Every man has his price" ' (ibid., p.202).

However, Caswell fails to explain what the 'disposition-aspect' of a meta-maxim is, and in particular how it contrasts with the 'propensity-aspect'. This may be because once he has characterized the propensity-aspect as the ground of an agent's deliberative bent, he has left himself little or no conceptual space for a distinct characterization of the disposition-aspect of the evil meta-maxim. Perhaps he could have said that the disposition-aspect consists in being the sort of person who is disposed to a susceptibility to evil choice. But there seems little or nothing to distinguish this from his characterization of the propensity-aspect. For these reasons, it does not seem as though we have grounds to take the notions of disposition and propensity as aspects of *a single entity*: i.e., of a meta-maxim.

In contrast, as we have seen, Morgan's account gives very real and distinct roles to the notions of a propensity to evil and the good and evil dispositions. The propensity to evil is understood as a temptation in the form of the incentive to license to embrace (individual evil maxims) or even an evil overarching principle. Those with evil dispositions have done the latter (i.e. they possess the evil meta-maxim). People with good dispositions have the moral meta-maxim but are still subject to the propensity to evil and may succumb to temptation and perhaps even embrace an evil overarching principle.

In addition, I find it difficult to make sense of Caswell's (and Kant's) idea of a propensity or susceptibility as *itself* a *ground* of choice. In contrast, in Morgan's reconstruction, the propensity to evil is seen as the presence of an *incentive* to license. This has the (technical) advantage that, in Kant, it is an incentive which if incorporated into a maxim issues in the adoption of that maxim, and this is indeed how the incentive to license functions in Morgan's reconstruction: if the incentive to license is endorsed at the highest level, the agent thereby adopts the evil meta-maxim. But in Morgan's account, the incentive *qua* incentive is not seen as a ground of choice. If the incentive is incorporated, then the adopted meta-maxim becomes that ground. Morgan is here applying the notions of incentive and maxim in accordance with the absolutely central Incorporation Thesis in his rendition of the propensity to evil understood as incentive to license. If Caswell's analysis is close or identical to Kant's position on this, then Morgan turns out to be more Kantian than Kant in this respect.

Furthermore, if Caswell's analysis sees the possession of the moral meta-maxim as moral perfection, then this is perhaps what gives his interpretation of the propensity to evil as an *aspect* of having an evil meta-maxim its intuitive point since it highlights an idea which he thinks might not be immediately obvious: that people who are generally good (but merely generally so) actually have an evil meta-maxim (and in some unclear sense an evil disposition), *contrary to expectation*, and thereby have a will with a deliberative orientation grounded in a susceptibility to evil. The point of this propensity-aspect of an evil meta-maxim would then be that those who generally adhere to the law need to be vigilant if they wish to become truly virtuous. But if he thinks these generally good people prioritize self-love over duty, is not the possibility of moral failure already captured by this prioritization? Why do we need to posit an aspectual ground of choice (the propensity) to capture this idea of fallibility?

Perhaps Caswell would say that the possibility of moral failure inherent in practical deliberation might be clear to the theorist who takes such generally good people to have adopted the evil meta-maxim but it might not be obvious to the agents themselves and part of Kant's project was actually to help people to become virtuous. So perhaps the point would be the message: until you are perfectly good, you are evil and this means moral fallibility is at the very root of your practical deliberation. However, if the conception of possessing a good meta-maxim as moral perfection is what gives his account of the propensity to evil its point, then Caswell faces a problem: we might think that this is an unintuitive conception of a good disposition (if disposition is understood as possessing a meta-maxim) since it entails that a good person is always good and cannot suffer a fall and this may be too high a price to pay for giving his analysis of the propensity to evil its point.

Unfortunately, at one point in his paper (ibid., pp.203-4), Caswell also seems to take the view that the will which prioritizes duty over self-love is *not* perfectly good. This view of the moral meta-maxim involves a departure, on Caswell's part, from the maxim-rigorism that he earlier took Kant to hold but he never explicitly acknowledges this. This revelation appears in the context of his discussion of the acquisition of virtue. Briefly, for Kant, virtue is 'the capacity as *strength* with which the virtuous agent may 'overcome ... opposing sensible impulses' and 'is something he must acquire' (MS 6:397). It is the ability to resist evil itself - that perverse temptation to take desire satisfaction as more valuable than duty. Virtue admits of degrees: it is the degree to which one values the moral law over the temptation to endorse and so, act on immoral desires.[49]

At this point in the paper, Caswell correctly notes that the task of becoming virtuous consists in the overturning of evil and that 'Kant describes this overturning of evil as requiring first

[49] I will not enter into how virtue is developed here since this will be dealt with in depth in Chapter 6.

and foremost a "change of heart" ' (Caswell, 2006a, p.203), which Caswell describes as a resolution 'to further ends of self-love only on the condition that moral interests are fulfilled' (ibid.). I think it is clear that here he means the prioritization of the incentive of duty over that of self-love: the adoption of the moral meta-maxim.

I think Caswell is right about this: for one thing, we can see that Kant's view is indeed that the revolution is the first step and not the last on the road to perfect virtue in one passage in the *Religion*. Here he says of a person who

> reverses the supreme ground of his maxims by which he was an evil human being (and thereby puts on a "new man")' that he thereby becomes 'receptive to the good' and that he can now hope by virtue of having adopted this principle 'to find himself upon the good (though narrow) path of constant *progress* from bad to better. (R 6:48)

However, if adopting the moral meta-maxim is only the first step towards moral perfection, then doing so cannot be sufficient for perfection itself. If Caswell's view *now* is that the adoption of the moral meta-maxim alone is not moral perfection, then he must admit that the agent who has only just adopted it is still subject to the propensity to evil. However, this seems to contradict his earlier analysis of the propensity to evil as an aspect of the adoption of the evil meta-maxim. *Then* he argued that the propensity to evil was a ground of practical deliberation *emanating from the possession of the evil meta-maxim*. In short, *now*, in his discussion of virtue, Caswell takes the same position as Morgan: that the adoption of the evil meta-maxim (which is for Morgan, equivalent to the evil disposition) and the propensity to evil are distinct because now he must say the propensity to evil persists even when the evil meta-maxim does not.

Caswell may object that after the revolution, the agent still has a propensity to evil but that this is no longer a *ground* of their deliberation. But to understand what he means by 'propensity to evil', all we have to go on is his characterization of it as *an aspect of the possession of the evil meta-maxim* and as a ground of deliberation. Even if we grant that it is no longer legislative after the revolution, it is not clear from this what it could be and where it emanates from. Conversely, Morgan's reconstruction allows us to see how any agent - whether their meta-maxim is good or evil - can come to be tempted by the incentive to license, since the incentive is ever-present, although Morgan's position is that a will with a good disposition is better fortified against this.

So Caswell must choose between the following: he can see the adoption of the good meta-maxim as perfection. This would give point to his arguments concerning the intuitiveness of ascribing an evil meta-maxim (with its 'aspect' of a propensity to evil understood as moral

fallibility grounding the practical deliberation) to generally good people. But then he would have an unintuitive conception of the moral meta-maxim (i.e. as sufficient for perfection); or he must agree that the propensity to evil must somehow persist in those who have adopted the moral meta-maxim but not yet attained perfect virtue. But as we saw above, his account offers no explanation for this since the only explicit account of the propensity to evil he has is one in which it is *bound* to the possession of an evil meta-maxim.

Morgan sees the revolution as the beginning of the process of the development of virtue, so for him, the adoption of the moral meta-maxim does not constitute moral perfection. This allows the intuitive idea that a generally good person is one who possesses the moral meta-maxim. But in separating the evil disposition and propensity to evil, Morgan can also take this good agent as still subject to the latter. This is also intuitive because it captures the idea that even for the best (though imperfect) the incentive to license may always present itself since, it is thought, the will may always misrepresent its freedom. How or why this might happen is perhaps the deepest puzzle in the practical philosophy. If this seems unsatisfactory, we should note that Morgan's reconstruction of evil has already gone further than Kant's account, since it elucidates how we might have an incentive to adopt the evil meta-maxim and (to my knowledge) Kant never attempted this.

Morgan's account contains a further insight - important to this study - into the nature of the propensity to evil and the struggle to develop a will which can resist it: the struggle for virtue. Recall that the challenge for the agent who has willed the pursuit of virtue is not just that empirical incentives are tempting and incorporating them sometimes goes against the law but involves a deeper struggle to reduce his propensity to see pleasure or happiness as more valuable than morality. This is not merely a propensity to overlook duty; it is an *active* tendency towards taking self-love as legislative. Caswell, like most orthodox commentators, brings out this idea to an extent but as we shall see, fails to provide the depth of analysis we see in Morgan's account.

Caswell says (quite rightly) that in Kant the adoption of any maxim involves the setting of an end and the meta-maxims are no different in this. This is why in Kant, 'Human reason has the peculiarity of framing for itself an ultimate subjective end in the indeterminate idea of happiness.' (Caswell, 2006a, p.205) He says that because this is an indeterminate idea, 'our ineliminable concern with happiness is already staggeringly ambitious.' (ibid.) The essential characteristic of the will with a moral meta-maxim is that it prioritizes duty over self-love. Caswell very reasonably argues that this is why the end of the good will is the highest good, 'conceived as maximal happiness in proportion to complete virtue.' (ibid., p.203) Conversely,

when we adopt the evil meta-maxim 'we will our own happiness unconditionally, and will virtue only in so far as it does not interfere with our own happiness.' (ibid.)

If we were to ask why this is the end of the evil will, then on Caswell's analysis, we can only appeal the propensity-aspect of the evil meta-maxim. We cannot appeal to his notion of the disposition-aspect of having an evil meta-maxim because Caswell has no developed account of this aspect. The propensity-aspect tells us that this will has the possibility of moral failure inherent in its practical deliberation. We could then say that where there is a moral failure, 'pressure', so to speak, which the end of indeterminate happiness exerts is allowed full rein and this how Caswell can capture the dynamic property of radical evil. But this would be a negative account which draws on the idea of the indeterminate happiness of any will, combined with a *failure* to value duty sufficiently. What is needed is a positive and richer understanding of why evil actively seeks to will 'worldly' maxims of action over those of duty.

I think Morgan's account offers such an account: when we (as theorists) see the incentive to license as involving the misrepresentation of freedom as consisting in the absence of all constraint and the boundless pursuit of *outer freedom*, we see that this will takes it that it can only affirm its spontaneity by adopting a principle which recommends the adoption of maxims of action with *empirical* ends - those which are endorsed to bring about happiness as the agent sees it. In addition, we can see why the evil will *actively resists* any return to a true conception of freedom because it is interested in expressing its spontaneity in the way it sees fit. This elucidates the way the struggle for virtue is a struggle against an active resistance to morality - or to put it another way - an active resistance against a true conception of freedom.

Once again, separating the propensity to evil and the evil disposition allows Morgan's theory several intuitive advantages over Caswell's analysis: since the propensity to license is not bound to the possession of an evil meta-maxim, it allows him to say that generally good people have the moral meta-maxim and are still subject to the propensity to evil: the incentive is an ever-present temptation to them. In addition, Morgan's conception of the moral meta-maxim is not one of perfection, so, again we can say of generally good people that they have adopted the moral meta-maxim and that this involves a defeasible commitment to good. Evil people have simply embraced the incentive to license at the level of the meta-maxim and therefore have the evil meta-maxim (i.e. a commitment to pursuing outer freedom licentiously).

3. Sussman and the 'anthropological' account of radical evil

As we have seen, Kant believes evil emanates from freedom itself and Morgan's account elucidates how this is possible. However, in his article, 'Perversity of the Heart', David Sussman offers a rival 'anthropological' reconstruction of the development of practical reason, moral norms and radical evil, mainly drawing on the *Religion* and *Speculative Beginning of Human History*.[50] This is thought of as an *anthropological* account because it sees the development of these as arising through a natural process. As we shall see, Sussman's account of the way this process is supposed to work is suspect. I will also argue that he is also unable to include the idea of a moral meta-maxim (as he wishes) in his account of the development of moral norms yet the inclusion of this seems to be required for the possibility of progress towards virtue. I also think there are difficulties in squaring any anthropological account of the development of practical reason, the norms of morality and the claims of self-love with some of the major claims Kant wants to make about these: i.e., that reason demands we abandon self-love as legislative and that he thinks we are responsible for our own evil.

Before we examine his account proper, we should first understand what Sussman takes evil fundamentally to be. In Section 1, I explained that Kant sees the propensity to evil as having three stages, the worst being depravity, then impurity and finally, frailty. Sussman wishes first to eliminate the 'less bad' stages (impurity and frailty) as the fundamental ground of evil, leaving depravity as the only remaining possibility for this within Kant's strictures. First, Sussman rejects any reduction of evil to weakness of will (i.e., frailty). He thinks one might be tempted to do this because in committing evil acts, an agent 'acts against norms implicit in her own autonomy, norms to which she is supposedly committed above all else.' (Sussman, 2005, p.155) However, he regards weakness of will as episodic and yet evil must encompass vices such as envy, resentment and bigotry, which can inform 'an enduring outlook on life' (ibid., p.156).

He also thinks 'impurity cannot be the fundamental ground of human corruption.' (ibid., p.160) Kant's impure agent has a moral meta-maxim but can only carry out moral acts (or at least acts in accordance with duty) with the help of inclination. The puzzle is why a fundamentally good person who, as such, (supposedly) takes the moral law to be supremely authoritative would 'have a standing need to represent their duties as being in their self-interest'. There must be some evil more fundamental than impurity which makes them value the pursuit of self-love to a degree which rivals duty, i.e., something which could bring about impurity.

---

[50] From now on, I will refer to this work as *Speculative Beginning*.

For these reasons Sussman takes radical evil fundamentally to be Kant's notion of depravity - the prioritization of self-love over duty. In order to explain the nature and origin of such evil, he deploys Kant's notion of the predispositions to good, which we met briefly earlier in this chapter.[51] These are the predispositions to animality, humanity and personality. He says, 'Each predisposition specifies a type of norm and a conception of oneself as an agent that is defined in terms of the ability to act in response to that norm.' (ibid., p.162) However, he thinks in attempting to realize each predisposition, we pervert it. This, he argues brings about a development in our practical rationality and a new, higher order set of norms associated with the next predisposition but we are also left with a residual competing propensity to act according to a perverted version of the old norms. Let us examine the account in detail.

Recall that the first predisposition is animality. The norm associated with this is our natural needs. The animalistic man's inclinations of 'mechanical self-love' move him to satisfy these needs. Sussman says, 'Such self-love is not rational, but it does anticipate reason, giving animal behavior enough implicit rational structure to be understood not just as movement, but as purposive striving.' (ibid., p.163) He then draws on *Speculative Beginning*, claiming this text sees progress from animality to the next predisposition, humanity, as made through perversions of the pursuit of animal needs. Animal man begins by determining what food is safe 'directly' through instinct and through those senses most closely associated with it: taste and smell. However, he begins to regard himself as one animal among others and to understand their activity as 'purposively directed toward the same general needs' (ibid., p.167) as his own. These realizations allow him to allow himself to try the foods they eat, thereby breaking from instinct. This, it seems, is the beginning of practical reason. It is also the beginning of the development of 'desires for things for which there is not only no natural urge, but even an urge to avoid' (MA 8:111) Sussman thinks that in wanting things he does not need, man has acquired a propensity to voluptuousness: the desire for pleasure itself rather than for things (which give pleasure). Drawing on the passage on the predispositions in the *Religion* (R 6:28), Sussman says that, at this stage, in addition to gluttony, we develop two other 'bestial vices': lust and wild lawlessness. He does not explain how these might come about. Perhaps the thought is that once man has acquired a taste for voluptuousness in one sphere of life (nutrition), we acquire it in others (sex and physical freedom).

Sussman thinks what is allowing these excesses is our falling foul of a *practical illusion* which consists in 'mistaking a subjective element in the grounds of action for something objective' (ApH 7:274; quoted in Sussman, 2005, p.167). Although Sussman does not spell it out, the role of this idea in the story seems to be that in taking the grounds of the pursuit of pleasure to be objective, we 'overdo' that pursuit: Sussman argues that the instincts for food, sex and

---

[51] See Section 1.

physical freedom, which stood man in good stead become perverted into the vices of gluttony, lust and wild lawlessness mentioned in the *Religion*. He explains how we can see how these are harmful to the needs which their previous uncorrupted versions were supposed to serve: e.g., the result is now obesity rather than healthful nourishment, onanism and incest rather than child rearing, and reckless physical risk-taking (perhaps rather than prudent or moderate risk-taking).

Sussman claims that we then inaugurate our norms of honour passing from animality to humanity through our sexuality. In *Speculative Beginning*, Kant says that when man clothed himself, he 'passed over from mere sensual to idealistic attractions, from mere animal desires eventually to love' (MA 8:113) Sussman thinks that Kant is claiming here that the desire for the more abstract love and admiration issued in the inception of a new set of norms - those of honour, which are associated with the predisposition to humanity. However, even if we grant that norms of honour would be brought about by these new, abstract desires for love and admiration, it is nevertheless hard to see what role the perversions of animality - lust, gluttony and wild lawlessness - play in our progress from animality to humanity. This is an important problem because his main thesis is that we progress to the next predisposition by perverting the previous one. Also, as we have seen, Sussman sees sexuality as the connection between these two predispositions but he does not explain what role a *perverse* sexuality - the vice of lust, with its onanism and incest - plays in the story. Presumably, it should somehow bring about the adoption of clothing, (since this, as we have seen, ultimately brings about the norms of honour) but Sussman is silent on this. Here is a guess: Sussman's account of the harmfulness of lust and the other two vices is immediately followed by his discussion of the passage from *Speculative Beginning* concerning our donning the fig leaf. We might infer from the fact that Sussman's discussion of lust precedes that of his discussion of the fig leaf that he believes that, in Kant's account, our covering ourselves was a *response* to an excessive and damaging preoccupation with sexual activity. If Sussman does believe this, then we can ascribe a role to sexual perversion in his explanation of our progress from animality to humanity. Unfortunately, interpreting Sussman as saying lust was followed by modesty relies on a very tenuous discoursal clue in his article, so we cannot give a definite role to sexual perversion.

In *Speculative Beginning*, the adoption of the fig leaf issues in a change from sensual to idealistic attractions, which leads to an appreciation of beauty. Apparently, Kant then rather abruptly switches to a description of 'decency'. His characterizes this as 'a propensity to influence others' respect for us by assuming good manners (by concealing whatever could arouse the low opinions of others)' (MA 8:113) in the same paragraph as the account of the emergence of an appreciation of beauty. This suggests a reading of Kant's argument that

diverges significantly from Sussman's. Kant's thought seems to me to be that clothing ourselves brings about a sense of decorum or seemliness which eventually becomes a general, all-pervasive concern to be seen to be honourable, which is not limited to its original sexual sphere. However, Kant's account for the adoption of clothing does not characterize it as a response to a harmful perversion of the animal norm of sexuality as Sussman claims.

Also, as I mentioned, there are two other vices associated with animality in the *Religion* (gluttony and wild lawlessness) but Sussman is unable to involve these in his *Speculative Beginning*-based explanation of the emergence of the norms of honour associated with humanity because Kant only accounts for this through sexuality in that work (but not a perverse sexuality as we have seen). Since Kant does not see the transition as being through, for example, nutrition in the *Speculative Beginning*, Sussman has no way to force the *Religion*'s vices of gluttony or wild lawlessness into the account of the transition. If Sussman is right and progress is made through perversion and the other two vices are relevant, we might wonder how *they* bring about the norms of honour according to Sussman.

Perhaps the role of perversity in our progress through the predispositions is to do with the way that a rudimentary practical reason emerges when our practical thought breaks away from the dictates of instinct and it is certainly true that without practical reason, the characteristic activities of humanity - and personality - would be impossible. But this 'incipient' practical reason emerges *before* the onset of voluptuousness, so the role the latter plays in the development of practical reason is still a mystery. Perhaps the interest man develops in pleasure itself - i.e., his voluptuousness - is supposed to increase the use he makes of his primitive practical reason, accelerating its development. Unfortunately, here I am reduced once again to guessing what can fill this gap in Sussman's reconstruction.

The role of perversity in Sussman's reconstruction is a little clearer when we move on to the transition from humanity to personality but is still problematic. The perversion of humanity is described as follows, 'The desire for love and esteem starts to distort itself because it presupposes a standard of self-worth that is purely comparative' (Sussman, 2005, p.169) There is no objective standard and 'Without any absolute norms to secure our sense of self-worth, we strive to avoid appearing badly off in comparison to others.' (ibid.) Sussman claims that there is at first jealousy and rivalry but then there is competition for honour so manic that we are brought 'to the point of being willing to disgrace ourselves in order to embarrass or humiliate someone else.' (ibid., p.170) We then have the perversion of humanity in the

form of the 'vices of culture' - envy, ingratitude and joy in others' misfortunes[52] mentioned in the passage on the predisposition to humanity in the *Religion* (R 6:27). However, Sussman thinks that through them emerges a fledgeling predisposition to personality: the envious in not wishing to be subordinate to others and the ungrateful in not wishing to be dependent on them both develop 'a nascent appreciation of our equal dignity as persons' (2005, p.170); whilst in feeling joy at the misfortunes of others, regardless of whether we have any personal connection with them, we lay the foundations for 'a kind of disinterested concern in the welfare of human beings simply as such' (ibid.) presumably since through such joy - malicious[53] though it is - we are at least no longer utterly contemptuous of them.

This story which Sussman invents to explain exactly how we might progress from humanity to personality through these vices is open to question. Kantian moral law is founded in *pure practical reason*, so the idea that the beginnings of an appreciation of an aspect of it: 'our equal dignity as persons' (ibid.) could be derived from a *disinclination* on the part of the envious to be subordinate to others or on the part of the ungrateful to be beholden to anyone is peculiar. Sussman owes us an explanation of how we get from *desiring* certain things to knowing we are commanded by *reason* to do certain things. Moreover, there is a similar problem with the argument from shameful joy: it seems bizarre to think that being motivated to care about the material welfare of others by the desire to be better or more honourable than them could develop into a positive 'kind of disinterested concern' for their welfare not least because the original concern is anything but disinterested (even if impersonal). Again, Sussman is proposing that this aspect of duty, something which in mature rational agents is grounded in pure practical reason, develops out of (a rather malign) inclination. And again, Kant has his own account in *Speculative Beginning* (MA 8:114) as to how our disinterested concern for the welfare of others comes about: in realizing our superiority to the other animals we realize our equality with other human beings as ends in ourselves. This may be just as poor as Sussman's perversion story but it does provide an alternative to it; it may be because of this and his eagerness to promote his reconstruction that Sussman fails even to acknowledge its existence.

Sussman's account, then, purports to explain how human beings acquire practical reason 'by progressively working through interstitial vices that mix features of a lower and a higher predisposition.' (Sussman, 2005, pp.170-171) It also seeks to explain the nature and origin of self-love:

---

[52] Sussman is unhappy with the translations of *Schadenfreude* he has encountered ('malice' and 'spitefulness') and thinks the German term sufficiently familiar to English speakers to use it in his text. I have used the English translation given in the Cambridge Edition of the *Religion*.

[53] Although, involving no personal malice.

> These vices emerge when a previously authoritative norm (health, honor) starts to be recast as a form of self-love relative to a newly emerging principle (honor, morality). Each vice involves a propensity to cling to some kind of self-love that is a vestige of the authoritative status that such self-love is losing. (ibid., p.171)

So, for example, the norm of animality is health but when a new predisposition (humanity) emerges and with it a new norm (honour), the previous norm (health) becomes self-love and we are therefore supposed to prioritize honour over health just as in personality where we are supposed to prioritize morality above the vainglorious pursuit of honour.

The reader may judge from the preceding discussion of Sussman's account that it concerns the epochal development of the practical reason of the species and one could be forgiven for such a judgement since the *Speculative Beginning*, upon which his account is based, is one of just this sort - at least, I take it to be such. However, Sussman actually proposes that his account concerns the development of practical reason in the latter-day individual, within her lifetime. One problem with this is that it is unclear at which stages of life this development is supposed to take place. We can quickly reject any suggestion that he regards the three predispositions as corresponding to the life-stages of childhood, adolescence and adulthood: the notion that, for example, individuals progress from childhood to adolescence as they do from animality to humanity, according to Sussman - through the bestial vices such as sexual perversion - would be preposterous. Another suggestion which must be rejected is that Sussman believes that adults have all three predispositions and these must be developed simultaneously: throughout the text he speaks of progress from the first to the second and from the second to the third. We must also reject the option that he is saying we are at a point in history in which we have reached the stage of humanity and so all latter-day individual adults begin at least at that stage and must endeavour to progress to personality. Something that might recommend this interpretation is that it fits Kant's claims that (*a*) we are uniformly evil (i.e. of the human type as Sussman sees it) and (*b*) that it is a duty to develop our capacity to be moral (i.e., according to Sussman, to become a person). The reason I reject this is that Sussman sees animality as the first stage, not humanity. The only option remaining is that he believes all adults progress through all three predispositions sequentially.

The most remarkable thing about Sussman's reconstruction is that although an essential feature of Kant's notion of predispositions, as they appear in the *Religion*, is that they are *original*, i.e., essential (and Sussman is well aware of this), he makes them something we *acquire* in his reconstruction. From an orthodox Kantian perspective, such a conception of personality would be problematic because if the predisposition to personality were not actually that - a *predisposition* - i.e., if it were not *original*, the agent would have nothing to

guide them in - and no motivation for - moral development: if we had no conception of morality, we would not be able to become more virtuous. So, an orthodox critic might say Sussman's idea of progress towards the good is impossible because progress itself requires some initial conception of morality, which neither the animal nor the human agent has.

However, Sussman's model is different because he sees our progress through the predispositions as driven by *a natural process* in which practical reason itself and its norms *emerge* through antagonisms with irrationality and are only realized at the stage of personality where we finally conceive of ourselves as rational agents *simpliciter* and thereby also recognize the norms of morality which go along with such agency. Since the process is a natural one, it by-passes the need to posit an original predisposition to personality (i.e. to morality) as a guiding and motivating incentive to pursue virtue. Instead, presumably, once nature has led us to personality - the state of being fully rational and therefore subject to morality - we recognize the moral law and it then becomes possible for us to begin to pursue virtue by increasing our identification with the self-concept of a rational being *simpliciter*, by focusing on our new, rational set of moral norms and by moving away from, as Sussman sees it, the irrationality of self-love associated with humanity.

However, Sussman creates a problem for himself in trying to incorporate Kant's idea of a noumenal choice of a meta-maxim into his account of our progress towards practical reason and morality. Sussman demystifies this notorious idea of a 'timeless' choice by arguing it is to be seen as timeless in the sense that a good person is one whose life *as a whole* consisted of an earnest striving towards personality (rather than something chosen at a particular time) whereas an evil person is one whose progress faltered at one of the vicious stages. However, it is hard to see how he can combine this conception of the choice of a meta-maxim with his account of progress through the predispositions, which we have been discussing. In order to be a good person, one has to make progress towards personality. In order to do that, one must adopt the moral meta-maxim. But to have any conception of that maxim, surely, one must have some conception of morality. However, according to Sussman, we all begin at the stage of animality and animal-man has no conception of morality, so how is such a choice possible for him? Sussman may say that it is not necessary to choose a meta-maxim in order to progress: instead if one made progress or at least earnestly tried to make progress in one's life, then it *turns out* that one's whole life consisted in a choice to pursue the good (which choice is tantamount to the possession of the moral meta-maxim). But we may wonder what role the idea of a supreme maxim is playing in all this. If it turns out at the end of his life that an agent was good 'overall', then when he was at the stage of animality, he would have been working towards the next predisposition - perhaps with some setbacks along the way - but the overall trend would have been upwards. But for the idea of a meta-

maxim to be doing any work in the theory, we must say that it had some influence on the agent's choices - choices which somehow issued in his progression to the next predisposition. However, if that is so, then we are positing an animalistic man who can respond to a meta-maxim through the hierarchy we met in Chapter 2. But this is an impossibility because *the animalistic man is not rational.* The problem is that the doctrine of *Gesinnung* and its attendant hierarchy of maxims are simply fundamentally unsuited to this sort of naturalistic explanation of the development of practical reason, moral norms and radical evil.

4. Conclusion

As we have seen, there are problems associated with the way interstitial vices are supposed ultimately to inaugurate the norms of morality in an individual in Sussman's account. In addition, he seems to claim that moral progress begins at the (non-rational) stage of animality and yet he wants to see such progress as taking place in the context of an adopted meta-maxim, which, as I see it, can only be adopted by a rational being. Aside from these difficulties with Sussman's account, it is not clear how any 'anthropological' account of the development of practical reason and the recasting of the norms of honour as self-love through a natural process can be can be squared with important features of the practical philosophy. For one thing, it is hard to see how we can hold on to the claim on behalf of Kant that reason *commands* us to abandon the legislative claims of self-love when according to such a theory we can leave them behind over time simply through the continual exercise of practical reason. Furthermore, if evil develops naturally, then we are not responsible for it, which suggests that we are not responsible for resisting it but this last claim was one that Kant thought necessary to make.

For these reasons I reject Sussman's reconstruction and its view of self-love: namely that we should think of the claims of self-love as claims which for some agents were once norms (of honour) but have been supplanted as such, for those agents, by the norms of morality through a natural process. The other options for an account of radical evil considered in this chapter were Morgan's rational reconstruction and Caswell's more orthodox analysis of Kant's account. I reject the latter as it offers none of the advantages for the former. Morgan's reconstruction provides a formal proof which also vindicates many of Kant's claims concerning the propensity to evil, as we have seen. It also gives real, distinct and I think intuitive roles to the idea of a propensity to evil and the evil disposition in that the propensity is seen as an inextirpable but combatable incentive or temptation to adopt a policy of unlimited license which the evil have endorsed at the meta-maximal level and the good may endorse. This avoids the potentially disastrous notion that in positing a universal and inextirpable propensity to evil and on the face of it identifying this with an evil *Gesinnung*, Kant seems to preclude a

moral revolution in one's supreme maxim. It also allows us to see a good disposition as a fundamental but defeasible commitment to morality, which is perhaps what we would expect a good person to have. Finally, in characterizing evil as something which emerges from the spontaneity of the will itself, it can be thought of as something which *strives* to continue to be the will's overarching policy once initially accepted. This it must do in the face of the will's overriding reason to choose morality, (since morality expresses its true freedom). In the next chapter, I will argue that the adoption of evil requires that the will deceive itself into supposing that freedom lies in license. Since the model of self-deception which I intend to deploy is one that takes self-deception to be *intentional*, Morgan's picture of evil as something which when it gets a grip on the will, is 'interested' in maintaining its position coheres well with this view of self-deception.

Chapter 4

Self-deception in the will and the adoption of evil

In Chapters 2 and 3, I alluded to the belief that self-deception is involved in Kantian radical evil. Presently, I will argue that the nature of that involvement is that self-deception on the part of the will itself is a condition of the possibility of evil. This is because the will which chooses evil must do so despite in some sense knowing that it has overriding reason (in its nature as freedom) to choose morality. Evil cannot be chosen by accident or in ignorance because it is something for which the will must be responsible. The only other way the choice can be facilitated is by self-deception with regard to freedom (so long as self-deception is construed in a way which makes the deceiver responsible for it). I argue that it is therefore required for a choice of an overarching policy of evil in the form of the evil meta-maxim and particular evil maxims. However, self-deception is *prima facie* paradoxical firstly in that it seems to involve the holding of contradictory beliefs (or in the case of the will, perhaps some analogue of belief). Let us call this the *Belief Paradox*. An account must be found which dissolves this paradox before the notion of self-deception can be included in the practical philosophy. Any self-deception story must also be able to dissolve what I wish to call the *Deceiver Paradox*:[54] that of a self-deceiver's both knowing (as deceiver) and not knowing (as the one deceived) about the process or the act of deception. This paradox must be dissolved (i.e., it must be shown how an agent can successfully engage in the process of self-deception) in a way which preserves agential responsibility before we can say that the self-deception story is suitable for inclusion in the account of a choice of evil in Kant.

In this study, additional requirements stem from the fact that the chosen account is to be incorporated into the practical philosophy and more specifically into Kant's doctrine of evil. One issue is that accounts of self-deception unsurprisingly deal with that affliction in, as it were, 'whole' human beings, yet the present study seeks to apply such an account to the rather more rarefied entity of the human will. A self-deception story cannot be deemed suitable if it demands that its subject be able to do things (e.g., have a capacity for belief) in order to achieve a state of self-deception which the Kantian will cannot do. We will see that the will has each capacity (or at least an analogue of it) required for self-deception. A further challenge which arises from the use of a human self-deception story is that obviously it will take the development and maintenance of self-deception to be something which occurs straightforwardly in time. The difficulty to be overcome is that such an account must somehow be applied to our picture of the acquisition of a meta-maxim. This acquisition (as

---

[54] This is short for perhaps the more apt name, "Deceiver-Deceived Paradox".

we saw in Chapter 2) whilst not actually timeless nevertheless does not take place at a particular point in time but rather one's whole life (or perhaps a period of it) is to be taken as amounting to that choice. I will argue that Sartre's account of bad faith from *Being and Nothingness* can dissolve the paradoxes and preserve responsibility thus providing a self-deception account well-suited to the Kantian will itself.

1. Evil and the paradoxes of self-deception

1.1 The evil will must be self-deceived

In Chapter 3,[55] we saw how the will must have a supreme maxim to give it a reason to choose (or reject) lower order maxims and form a hierarchy and how, according to Morgan and Korsgaard, a free will has overriding reason to choose morality as its highest maxim. Since the maxim being chosen is supreme, the will has no recourse to a further principle for guidance in its choice. It finds a reason in its very nature as freedom to choose the maxim that best preserves this freedom - the moral meta-maxim. Making this claim commits us to two further claims: (*1*) that the will is capable of knowledge or perhaps understanding (or some analogue of one of these) and (*2*) that the will knows that it is freedom itself. I think we can safely make these two further claims for the following reasons: regarding (*1*), that the will is practical reason is perhaps reason enough to suppose that it can, in some sense, know. But in addition to this, some of the most important features of the practical philosophy depend on the its being able to do so. For example, it could not be a faculty of choice if it could not know what it was choosing. Nor could it make use of its higher order maxims to guide it in its choice of lower order ones if it were not in some way aware of what they recommended. Finally, some form of knowledge capability must be ascribed to the will if the agent is to be held responsible for the choices of maxims he makes through it. As regards (*2*), if the will did not know that (as practical reason), it is free - that it has the property of autonomy - any will that endorsed the incentive to license and thereby chose the evil meta-maxim could not be held responsible for doing so, nor could Kant posit a predisposition to good in the human being.

However, this presents us with a puzzle: given that every will knows its own nature as freedom consists in autonomy and so knows that it thereby has overriding reason to choose morality, it is difficult to see how or why any will would endorse the incentive to license and choose the evil meta-maxim. Although it was established in the last chapter that the practical philosophy need not be constrained to posit *universal* evil (in the sense of the universal possession of the evil meta-maxim), that there are people in the world who could be classed

[55] Sub-section 2.1

as evil in Kantian terms is nevertheless clear since, unlike moral action, the moral status of many evil acts is unambiguous and there are those who spend their lives carrying out such acts. Such people could not have adopted the moral meta-maxim, so following rigorism, we must say they have adopted the evil one. The puzzle then is *how a will which knows its essence as freedom is only preserved by the moral meta-maxim accepts license as freedom and thereby chooses the evil meta-maxim.*

The notion that a choice of evil could be the result of ignorance on the part of the will that chooses it must be rejected. If this is to count as a choice of *evil*, it must be something for which the will is responsible. This in turn requires that the relevant act which brings about the policy of evil is genuinely a *choice*. But *qua* choice, it must be an act which is carried out by one that understands the options involved. For the same reason, we must also reject any notion that accepting license and thereby choosing the evil meta-maxim could be an accident (at least it could not be one which the agent could not be held responsible for guarding against). If the 'evil' maxim were adopted in ignorance or by accident, the agent would be exonerated from any wrong-doing both for having that policy and for the individual acts stemming from it. Those acts would no longer be evil. If the will cannot be ignorant of its overriding reason to choose morality and if a choice of evil cannot be an accident, the only way to explain how it can knowingly choose evil seems to be to say that the evil will (in some sense) deceives itself with regard to freedom.

1.2 The Belief Paradox

Let us now turn to the paradoxes of self-deception. Although it is perhaps not necessary to talk in terms of propositions and propositional attitudes in order to address self-deception, much of the literature discusses an agent's self-deception with regard to some proposition, *p* and the present exposition of the paradoxes may be clearer if explained in propositional terms, at least for now. One paradox whose importance is emphasized in the literature - what I have dubbed the *Belief Paradox* - is that (at least *prima facie*) *a self-deceiver both believes p and believes not-p at the same time.* In the case of the Kantian will, the relevant true proposition, *p*, might be something like, 'My nature as freedom consists in autonomy.' The Kantian will knows this to be true and yet somehow deceives itself with regard to it, wilfully to go against its own nature and choose evil.

1.3 The Deceiver Paradox

As we have just seen, evil is a free choice in Kant (and arguably must be so since otherwise it is hard to see how it could be regarded as evil). It is therefore something for which the agent is responsible. I have argued that self-deception is a condition of the possibility of such a choice.[56] Given this, the will must *also* be responsible for its own *deception* since if it is not, then a 'choice' of 'evil' facilitated by a 'self-deception' for which it was not responsible would neither be a choice nor evil having been facilitated by some debilitating state[57] it was in, through no fault of its own. This would make the incorporation of a 'self-deception' story self-defeating since its purpose is to *preserve* responsibility for a choice of evil (as I argued in Sub-section 1.1). This responsibility requirement generates the potential for the *Deceiver Paradox* since if self-deception is an action or a state brought about by one's own actions and through one's own choice, then it seems *the self-deceiver must both be both aware (as deceiver) and unaware (as deceived) of the process of deception.* He must be unaware because obviously one cannot be deceived if one knows about the putative deception but he must also be aware of the deception in order to execute it. This paradox is mentioned, for example, by Sartre (1957, p.49). In addition, Kant himself seems to have this particular paradox in mind when (in the context of explaining why lying in general, including lying to oneself, as he puts it, is wrong) he admits that,

> It is easy to show that the human being is actually guilty of many inner lies, but it seems more difficult to explain how they are possible; for a lie requires a second person whom one intends to deceive, whereas to deceive oneself on purpose seems to contain a contradiction. (MS 6:430)

If Kant's doctrine of evil requires that the will be self-deceived with regard to freedom and if self-deception involves the Belief Paradox and the Deceiver Paradox, then what is being proposed as a solution to the problem of how a will chooses a policy of evil is itself problematic. It should be noted that whilst this study is concerned with the relatively narrow issue of moral self-development, what is at stake here is much more important than that: if we cannot find a self-deception story which successfully dissolves these paradoxes and in a way which is suitable for incorporation into the practical philosophy, *the very notion of Kantian evil would seem to be impossible.*

---

[56] Since, as we saw in Chapter 3, the will has overriding reason to choose morality.

[57] I refrain from calling this 'debilitating state' self-deception since I take self-deception to be a state brought about by the agent himself and the debilitating state described is not. It would be, at best, some form of quasi-self-deception since, although it would not strictly be *self*-deception, it would have the same effects as actual self-deception (such as the agent's sincerely avowing what he knows to be false etc.).

2. Sartre and bad faith

2.1 Interpreting Sartrean bad faith

In this section, I would like to outline an interpretation of Sartre's notion of bad faith[58] with a view to applying it to the account of the Kantian will's choice of the evil meta-maxim. I believe this account dissolves the two paradoxes of self-deception outlined above whilst preserving agential responsibility. Crucially it does all of this whilst also positing an agent who intentionally deceives themselves. This is important because the will must do this intentionally (like anything else it does). Let us turn to the account.

Sartre believes that bad faith is a project whose aim is to alleviate a certain type of anxiety about our freedom. According to him, each of us chooses a certain set of projects. Our character traits consist in these projects and give rise to our seeing the world as making certain demands upon us. The anxiety we feel concerns these apparent demands and our responses to them. It is because the apparent demands are based on projects which are freely chosen that the anxiety we feel about them and our responses to them amount to an anxiety about our freedom. Since we are all free, this anxiety is a feeling universally felt and bad faith, a project universally undertaken.

Before we look at what bad faith is, it is necessary to explain some terms whose use is peculiar to Sartre. In his philosophy, there are two aspects of human existence: *facticity* and *transcendence*. Facticity seems to include one's body, environment and the history of one's freedom. Jonathan Webber (2009, p.76) argues that it also should be taken to include one's present character, which is determined by the projects one has freely chosen. Transcendence seems to be the power to change one's character - i.e., one's projects. In addition, Sartre posits two modes of being: being-in-itself and being-for-itself. Those things which have the former type of being - such as inanimate objects - are incapable of changing their properties. Those which have the latter type of being - e.g., human beings - are capable of change.

Sartrean bad faith is often taken to consist in taking one's transcendence as facticity or one's facticity as transcendence. Sartre's example of the café waiter is sometimes thought to illustrate the former. As a human being, the waiter exists in the mode of being-for-himself, so the trait of waiterliness is something he is capable of changing. However, his being a waiter and having waiterly traits being the result of a free choice is the source of some anxiety for him. His solution is to behave in a way intended to convince himself that he exists in the mode of being-in-itself and that his waiterliness is a fixed property: he exaggerates

---

[58] I thank Seiriol Morgan for the suggestion (made in personal correspondence) that Sartrean bad faith might be an account of self-deception suitable for the purpose mentioned here.

waiterliness to the point of caricature, moving in a hurried, mechanistic fashion and precariously balancing his tray as he weaves between people. By acting as though he is a waiter in the mode of being-in-itself, he can pretend he is unfree and thereby address the cause of his anxiety.

As mentioned above, one could argue that Sartre also proposes the opposite strategy of taking one's facticity as transcendence as a form of bad faith. This is meant to be illustrated by Sartre's example of the unhappy homosexual. This man looks upon his facticity - his past acts, which were in fact homosexual acts - as explicable in different terms: he claims, for example, to be a heterosexual with 'a certain conception of the beautiful which women cannot satisfy' (Sartre, 1957, p.63). What allows him to make such excuses is that he is able to regard himself as having being-for-itself and therefore to suppose that he cannot be defined by his actions. In taking this attitude, I place myself 'on a plane where no reproach can touch me since what I really am is my transcendence. I flee from myself, I leave my tattered garment in the hands of the fault-finder.' (ibid., p.57) (In this case, the fault-finder is himself.)

A woman on a date may seem to illustrate both strategies. Her companion showers her with compliments which could be interpreted as either respectful and non-sexual or as part of a seduction attempt. Her complex requirement of his desire is that firstly it address 'her full freedom' since 'the desire cruel and naked would humiliate and horrify her.' (ibid., p.55) And yet she also wishes that his desire 'address itself to her body as object' since 'she would find no charm in a respect which would only be respect.' (ibid.) She exhibits the first strategy of bad faith (seeing transcendence as facticity) in the following way: whilst she knows that his apparently respectful attitude towards her can transcend itself and be (or perhaps become) something else (part of a seduction strategy) she chooses to ignore this and arrest it in its facticity - its present, apparent meaning. Thus she is said to employ the first strategy of bad faith (of seeing transcendence as facticity). Some might say she exhibits the second strategy of bad faith (facticity as transcendence) when her companion takes her hand. To withdraw it would kill the moment but to simply leave it there would be detrimental to her self-esteem. To address this dilemma she leaves her hand in his *but also* 'speaks of Life, of her life, she shows herself in her essential aspect - a personality, a consciousness.' (ibid., p.56) Some might argue that in doing so, she is focusing on her being as transcendence in order to take her attention away from her being as facticity - in particular, her being as a body - so that her hand can remain in his, since she, as transcendence, is now divorced from it, as body.

Jonathan Webber has identified certain ways in which this interpretation of bad faith as consisting in this pair of strategies is incoherent and has proposed an alternative reading

which purports to avoid these problems. Perhaps the most important difficulty lies in the fact that bad faith is meant to be a response to anxiety over *freedom*. If that is the case, then whilst we can see why an agent might want to see their transcendence as facticity, it is difficult to see why they would want to adopt the opposite strategy of taking their facticity to be transcendence. Transcendence - the ability to go beyond one's situation - requires and involves freedom, so a denial of that freedom would address anxiety and so the first strategy makes sense. But it is not clear how bad faith could ever involve the second strategy - taking one's facticity to be transcendence - since given that the problem was anxiety over freedom, the project would not only be self-defeating but would actually exacerbate it. Another difficulty which Webber points out is that Sartre thinks that belief in psychological determinism is the basis of all attitudes of excuse. If bad faith is to be consistent with this notion, then all forms of it would have to involve the agent's taking himself to have fixed properties which causally determine his behaviour. This would rule out a form of bad faith which purportedly involves taking one's facticity to be transcendence.

It is for these reasons that Webber's interpretation of the types of bad faith all involve the agent's taking himself to have fixed properties. Let us turn to the taxonomy of types of bad faith as he sees it. He claims that Sartre uses the term 'bad faith' in a general sense which includes two main types of bad faith. The first type is called sincerity and involves taking properties which one actually possesses to be fixed. The waiter is an example of this since he exaggerates his waiterly properties to convince himself that he is a waiter in the way that the coffee machine is a coffee machine, rather than seeing himself as a person who *is* indeed a waiter but in a way such that he could be something other than a waiter. Webber calls the second type of bad faith in the general sense, 'bad faith in the narrow sense'[59] and this involves denying properties one actually possesses. This is further subdivided according to the two slightly different strategies which are employed. The first strategy is to deny one's actual properties by pretending to have different ones which are taken to be fixed. An example of this is the unhappy homosexual who denies his homosexuality by claiming the relevant actions are those of one with the (fixed) property of adventurous heterosexuality.[60]

---

[59] Webber argues that because Sartre says that the waiter is an example of sincerity, some have supposed that he is not an example of bad faith. Webber thinks this is wrong and that the reason for the confusion is the existence of the two senses of 'bad faith' we have seen so far: whilst it is true to say that the waiter is not in bad faith in the narrow sense, it is wrong to think he is not in bad faith in the general sense since sincerity is a type of bad faith in this latter sense and he exhibits sincerity.

[60] One who takes the view that the homosexual's bad faith consists in taking his facticity to be transcendence may point to the fact that in the example the man 'has an obscure feeling that an homosexual is not an homosexual as this is table is a table' (Sartre, 1957, p.64). Webber acknowledges that the man is aware of his freedom (to be something other than a homosexual), therefore, but denies that we must regard this awareness as intensional and can instead regard it as extensional: i.e., the man is aware *of* his freedom but is not thereby aware *that* he is free. In

The second strategy is to deny one's actual properties by emphasising other properties which one does in fact posses and again, these are taken to be fixed. Webber thinks Sartre's example of a woman on a date illustrates this type. When her companion takes her hand, she denies the property of her sexuality by emphasizing (what she takes to be fixed) properties of sentimentality and intellect.

Webber's reading of all of these examples overcomes the two difficulties which beset (what is perhaps) the more orthodox reading outlined earlier. On his interpretation, the agent regards themselves (or others) as having fixed properties in every case. This means bad faith can always (at least attempt to) address anxiety over one's freely chosen projects and the resulting demands the world makes of one.[61] In addition, in all of these cases there is no longer any clash with Sartre's notion that belief in psychological determinism is the basis of all attitudes of excuse. Also, as Webber points out, his interpretation allows Sartre to avoid certain difficulties peculiar to the example of the young coquette which the rival interpretation of bad faith outlined earlier fails to avoid. If the thought in the rival reading is that she first takes her companion's transcendence to be facticity and then takes her own facticity to be transcendence, then not only will the second strategy fail to address anxiety over transcendence but it will also be undermined by the first; as Webber asks, 'If I consider myself to be nothing but a free transcendence, how can I consider someone else not to be such without being acutely aware of the inconsistency?' (2009, p.85) To sum up, I endorse Webber's view that Sartrean bad faith involves taking oneself to have fixed properties, (whereas in reality one is free to change one's projects and thereby one's properties) as well as his taxonomy of the three types of bad faith in the general sense.

2.2 Bad faith and the dissolution of the Belief Paradox and the Deceiver Paradox

I think Sartrean bad faith dissolves the Belief Paradox by presenting us with something which plausibly counts as self-deception but which does not posit an agent who both believes the true but unpleasant fact and its agreeable negation - at least not one who straightforwardly does this. Instead, Sartre's agent knows the true and unpleasant notion. He says, 'I must know in my capacity as deceiver the truth which is hidden from me in my capacity as the one deceived. Better yet I must know the truth very exactly *in order* to conceal it.' (1957, p.49) In

---

short his awareness of his freedom does not constrain us to take this example as showing that bad faith can consist in taking one's facticity to be transcendence.

[61] Webber's analysis of the waiter example is essentially the same as the orthodox analysis of it since both claim that the waiter sees his waiterly qualities as fixed properties. It is only where the orthodox analysis claims a given example is one of an agent's taking facticity as transcendence (e.g., the unhappy homosexual) that Webber's analysis of that example differs since, according to him, all cases are examples of agents taking their properties to be fixed.

the case of Sartre's agent, this notion is that her character and the demands the world appears to make of her are the result of her freely chosen projects.

However, the paradox is avoided because, I would argue, it is not the case that the agent also (straightforwardly) *believes* the opposite:[62] i.e., it is not the case that she *believes* that it is not the case that the demands the world appears to make of her are the result of her freely chosen projects. If she actually fully believed this, arguably, she would have no cause for anxiety which she does in fact have. Instead of actually *believing* the false but pleasant notion, the agent pretends that it is true by distracting herself from the true but unpleasant thought by wilfully misinterpreting evidence - by pretending that it shows what she wants it to show. In the case of the coquette, we can see this idea of pretence and the deliberate misinterpretation of her companion's words:

> If he says to her "I find you so attractive" she disarms this phrase of its sexual background; she attaches to the conversation and to the behavior of the speaker, the immediate meanings which she *imagines* as objective qualities. (ibid., p.55; emphasis added)

The idea of pretence or make-believe is more clearly evident in the language Sartre uses to describe what the waiter is doing. 'He is playing, he is amusing himself' Sartre tells us (ibid., p.59). A little later he says of being a waiter, 'It is a "representation" for others and for myself' and 'I can only play at being him, that is imagine to myself that I am he' (ibid.). Finally, Sartre says by being the waiter 'as the actor is Hamlet, by mechanically making the typical gestures of my state . . . . What I attempt to realize is a being-in-itself of the café waiter.' (ibid.)

So although commentators may refer to an agent's bad faith in regard to certain false notions as *beliefs*, I think we ought not to take them as such. I regard the following passage as one in which Sartre explains (or attempts to explain) the attenuated sense in which bad faith is belief (if it is to be regarded as belief at all):

> The true problem of bad faith stems evidently from the fact that bad faith is *faith*. It can not be either a cynical lie or certainty - if certainty is the intuitive possession of the object. But if we take belief as meaning the adherence of being to its object *when the object is not given or is given indistinctly*, then bad faith is belief (ibid., p.67; emphasis added)

This is hardly a conception of belief in which an agent sincerely bases what they take to be true on evidence they sincerely take to be clear and unambiguous. This is not to say that bad

---

[62] At least, there is nothing, as far as I can tell, constraining Sartre or his commentators to claim that the agent must *believe* the false but pleasant notion (i.e., that they have a fixed nature) even though they use the words 'belief' and 'believe' in respect of an agent's epistemic relation to the false notion.

faith notions can be held with unwarranted certitude - they can - but they are not the same sort of thing as a belief in the sense just described, since, for one thing, they involve and require a capacity for fiction, which belief in a more orthodox sense does not. This brings us to the issue of the *project* of bad faith and the particular approach to evidence which it involves. Sartrean bad faith may dissolve the Belief Paradox as I have argued above but this will have been in vain if Sartre's account of the *process* of bad faith fails to deal adequately with the Deceiver Paradox since in that case the agent will be unable to be in bad faith.

The project of bad faith, as we have seen above, can involve taking evidence for one thing as evidence for another - e.g., the woman on a date who takes what are really the opening gambits in a seduction attempt as polite and respectful compliments. Her companion's words do not constitute persuasive evidence in support of her bad faith 'belief' that he has the fixed property of innocent respectfulness and she knows (and so is dimly aware that) the opposite is the case. Given this, it is hard to see how it is psychologically possible to treat the evidence in this way. Webber tells us that part of the explanation for this is that the agent exploits the fact that any evidence underdetermines belief and so any given item of evidence can be taken to support what one wants it to support. Returning to the example, whilst the man's words can and ought to be taken to support the belief that he merely wants sex, since this conclusion is not strictly required by the evidence, it leaves open the opportunity to suppose that her own (non-threatening) conclusion is warranted.

The trouble with the ploy of exploiting the underdetermination of belief and thereby accepting unpersuasive evidence is that it is a poor approach to reasoning. Since some agents may be dimly aware of this, it is an approach which is *itself* unpersuasive. Yet this is something which Sartre insists is intentional: it is a project. He says, 'One does not undergo his bad faith; one is not infected with it; it is not a state. But consciousness affects itself with bad faith. There must be an original intention and a project of bad faith' (ibid., p.49) Sartre simply seems to think that its being intentional is just part of what bad faith is. The notion of intentionality is particularly clear in cases in which the agent actually *manufactures* the evidence - such as that of the waiter. The waiter's act is a ploy to convince himself (and the others in the café) of his fixed waiterly properties. Now, whether it is a case of the person interpreting evidence which presents itself to them (e.g., the young coquette) or whether the person is interpreting evidence which they themselves have manufactured (e.g., the waiter), we need to know how the person in bad faith is pursuing a project of bad faith intentionally which involves intentionally having a bad faith attitude to the evidence *and a bad faith attitude to employing this method.* We need to know how this can be done without the account's falling foul of the Deceiver Paradox.

As Webber tells us, (2009, p.100), the first point to note is that it is possible to pursue an aim without one being fully conscious of it. An example makes this vivid: he says a person need not think of the aim of walking to work in order to pursue it. I think this is similar or perhaps reduces to a point he makes earlier in the chapter (ibid., p.92) that one need not consciously think of the aim of walking to work for that aim to unify a set of lesser actions (walking, avoiding obstacles, etc.) into that single overarching project (walking to work). Webber makes the further point (ibid., p.100) that whilst a person need not think of the aim, it can easily occur to him. Obviously, this would be fatal to a project of bad faith. However, he says that the individual in bad faith has good reason not to think about his project explicitly.

Webber also points out that in Sartre a project can structure my awareness not just of the world but also 'of myself - my experiences, my actions and my projects.' (ibid.) Not all projects will affect how I see myself - a project of walking to work may only affect how I see the world, e.g., as useful or as containing obstacles and so on. But in a project of bad faith, my actions, thoughts etc. will seem to manifest a fixed nature. In addition, the experience of all these seeming to emanate from a fixed nature itself seems to emanate from a fixed nature. This, he thinks, will obscure the relation between these things seeming this way and my project of bad faith. Presumably this is because my apparent fixed nature can 'take the blame' so to speak in place of the real 'culprit' - my project of bad faith.

We might wonder what it is about the project of bad faith which gives it the ability to do this. Webber points to Sartre's likening being in bad faith to being in a dream. Real events cannot penetrate the dream *as real events* because 'the real world is no part of the dream' (ibid., p.101):[63] the alarm clock sounding in real life must appear 'as the sound of drums or a fountain or even an alarm clock in the dream world.' (ibid.) In similar fashion, the experience of one's apparently fixed traits which actually emanates from the project of bad faith not only can but must appear to emanate from one's fixed nature because one is in the quasi-dream state of bad faith - within the dream it cannot appear to be what it is.

I think that Sartre's taking the dream-like state of bad faith to be what allows bad faith to make itself psychologically acceptable to the agent is suggested by what he says in the final section of the chapter 'Bad Faith',

> Thus bad faith in its primitive project and in its coming into the world decides on the exact nature of its requirements. It stands forth in the firm resolution *not to demand too much*, to count itself

---

[63] Webber tells us (2009, p.101) that the idea that real events as such cannot be part of a dream appears in another of Sartre's works: *The Imaginary*, where he develops 'a theory of dreaming . . . as part of his overarching theory of all forms of imaginitive consciousness.'

> satisfied when it is barely persuaded, to force itself in decisions to adhere to uncertain truths. This original project of bad faith is a decision in bad faith on the nature of faith (1957, p.68)

Here we clearly see the questionable epistemic attitude Webber said was part of the project of bad faith. This is closely followed by the following statements:

> One *puts oneself* in bad faith as one goes to sleep and one is in bad faith as one dreams. Once this mode of being has been realized, it is as difficult to get out of it as to wake oneself up; bad faith is a type of being in the world, like waking or dreaming, which by itself tends to perpetuate itself (ibid.)

Webber argues that given that bad faith is a mode of being such that it requires 'some cause external to the dream itself' (2009, p.101) to 'awaken' the agent, analogously everything in the dream will seem to confirm that everyone has fixed natures (the primary goal of bad faith). But crucially with regard to the present issue - the dissolution of the Deceiver Paradox - within the context of the dream, even those thoughts which constitute the pursuit of bad faith will seem to emanate from one's fixed nature rather than from the project as they in fact do.

3. Can Sartrean bad faith be incorporated into the practical philosophy?

Sartre's account obviously deals with self-deception in 'whole' human beings, as it were, but what is being proposed is an application of the account to the human will itself rather than the human being as a whole. We must therefore ascertain whether the Kantian will is capable of doing the things which the account requires the human agent to do to deceive himself. These things include the ability to know inexplicitly or dimly that freedom is autonomy, to believe in some attenuated sense that freedom is license, and to affirm a desire to take it as such. There must also be something which the will can take as 'evidence' to 'support' its self-deceptive claim and it must be able to do all of this without becoming aware of its own deception, despite carrying it out intentionally.

Firstly, to be in bad faith, the will must have a knowledge capability or some analogue of it since knowing the unwanted truth (as opposed to, say, merely suspecting it) is part of this model of self-deception. Clearly the will has this capability as it knows that it is freedom (as we established in Sub-section 1.1). Without this knowledge, the problem of self-deception would not even arise for it. However, Sartre's theory requires that this knowledge be concealed or made inexplicit. This is not an issue Kant explicitly addresses in any of the published works. However, there seems to be nothing in Kant's philosophy that would

preclude it and, moreover, there are certain features of the practical philosophy that suggest it is possible, perhaps even necessary in some cases. For example, we have seen that the will must in some sense know what maxims it has chosen. However, the following considerations might suggest that this consciousness of what is known might be 'dim' as it were, (i.e., not foregrounded but potentially available to full consciousness through reflection). First, it seems unlikely that all of the maxims the will chooses will be in harmony with one another. It does not seem outlandish to suggest that we can sometimes glean a person's intentions (perhaps over a long period of time) and the fact that those intentions are sometimes incompatible. So, a typically fastidious will, will have at least a few inconsistent maxims. But if it were the case that the will's knowledge of its maxims always had to be fully conscious (explicit and foregrounded), then (being practical *reason*) it would not tolerate any inconsistencies: arguably to consciously undermine one intention with another means that one or both of them are not genuinely willed. Finally, the fact (if it is a fact) that some inconsistency occurs, shows that the will is not fully conscious of all of its maxims - they are not all foregrounded - even though it must know them all. In short, if the will knows all of its maxims and if it always addresses any inconsistencies amongst them of which it is fully conscious, then if there are in fact some inconsistent maxims, then its knowledge of (at least some of) them must be inexplicit. The only other possibilities are to claim that there are no inconsistencies or that it consciously tolerates inconsistencies or that it does not know all its maxims. These all seem less plausible than ascribing dim or inexplicit knowledge or some analogue of it to the will. So, the fact that our candidate self-deception account takes this sort of knowledge of the unwanted proposition to part-constitute self-deception is no obstacle to the incorporation of that account to the practical philosophy. When we apply this part of the account to Kant, we say that the self-deceived will inexplicitly or dimly knows that its true nature as freedom can only be expressed through autonomy.

According to the theory, the self-deceiver must be able to believe, in some attenuated sense, the false but congenial notion - it must be capable of the faith of bad faith. Borrowing Kent Bach's[64] terminology, what seems to be required is the capacity to think *of* not-*p*, in order to avoid thinking *that p*, where *p* is a true but unwanted proposition. Unsurprisingly, there is nothing in the corpus to suggest either that the will can or cannot do this. However, if the will as practical reason is capable of knowledge or knowledge analogue as I have argued and is therefore capable of belief, there seems to be nothing about the will which would prevent its being capable of this sort of distracting thought. Applying this part of the account of self-deception to the practical philosophy, we say that the self-deceived will consciously avows or *thinks that* its nature as freedom consists in the unrestrained pursuit of outer freedom.

---

[64] These were notions he used in his papers, 'An Analysis of Self-Deception' (1981) and 'More on Self-Deception: Reply to Hellman' (1985).

Sartrean bad faith is desire-driven: Sartre regards bad faith as a strategy the agent employs because of an *aversion* to feelings of anxiety. This means that if we are to take Sartrean bad faith to be the model of self-deception through which the will chooses evil, we must say that in this particular case, the will is self-deceived with regard to its nature as freedom because it *desires* to take its freedom as license - i.e., it desires to engage in a process or project of self-deception. However, in order to be a desire which is actually acted upon - one which motivates the adoption of a maxim of bringing about bad faith - it must be a desire *accepted* by the will. But since this is a desire which motivates a maxim of self-deception which ultimately aims at the adoption of the evil meta-maxim, affirming that desire must itself be an evil act. Now, generally it is possible for a will to endorse evil desires without having to be an evil will but the desire in question is *foundational* to an *overarching* policy of evil. It is therefore hard to see how a will could affirm this desire unless it was already evil. At the beginning of this chapter, I argued that the will's being self-deceived with regard to its nature as freedom is a condition of the possibility of its choosing evil. It now emerges that so long as we use a self-deception story that requires the affirmation of an initial evil desire (as the account under consideration does), it is *also* the case that the will's choosing evil is a condition of the possibility its being self-deceived with regard to its nature as freedom.

Given that evil requires self-deception and self-deception requires (the affirmation of the evil desire which) requires an evil overarching policy, perhaps the only potential recourse available if we are to save Kant's doctrine of evil and if we are using a desire-initiated conception of self-deception is to say that the affirmation of the evil desire, the state of self-deception and the adoption of the evil meta-maxim are all 'equiprimordial'.[65] Recall from Chapter 3 how Korsgaard uses a conceit in which the will 'chooses' its meta-maxim 'before entering the world' to explain the conceptual priority involved in this act in terms of an analogous temporal priority (because it makes a difficult notion easier to understand). We may now extend this conceit to include the elements of self-deception: the will, instead of merely (*1*) choosing the evil meta-maxim 'before it enters the world', also (*2*) affirms a desire to take license as freedom (*3*) takes license as freedom and ignores its autonomy, thereby entering into self-deception with regard to its freedom, all 'at the same time'.

However, whilst we can see that once the will is evil, it can affirm the desire to be self-deceived with regard to its freedom and that once it is self-deceived with regard to its freedom, it can adopt evil, it is still unclear why the will would accept this 'package' of the

---

[65] I thank Seiriol Morgan for the suggestion (made in personal correspondence) that the solution to this problem might be the notion of an 'equiprimordial' choice. In Section 4 of this chapter, I explain how I think this notion can be combined with the view (expounded in Chapter 2) of the choice of a meta-maxim happening at no particular time.

affirmation of the evil desire, self-deception and the adoption of the evil meta-maxim. It is perhaps somewhat like the person who knows he becomes violent when drunk and knows that drunkenness prevents him from caring about his violence, deciding, when sober, to take a drink large and strong enough to intoxicate him. We must accept that this sort of choice is unintelligible. However, against any charge that this result is unsatisfactory, it must be pointed out that Kant himself thought that the choice of evil could not be understood: when discussing the possibility of the revolution in the *Religion*, he says, 'the fall from good into evil (if we seriously consider that evil originates from freedom) is no more comprehensible than the ascent from evil back into good' (R 6:45). Presumably it is comments such as this which Morgan is thinking of when he points out (2005, p.89) that since the will has an overriding reason to choose morality, a 'reason' to choose anything else is no reason at all. This means we, as theorists, have a principled reason to believe that any such choice *must* be ultimately unintelligible. Referring to Kant's point (made in the *Religion*) (R 6:21, 32, 39-40) Morgan says, 'To explain evil is to explain it away and this places it in the causal order. Hence to some extent, evil must remain a mystery.' (2005, p.89) It is nonetheless possible to illuminate it more than Kant did since we can illuminate *the sort* of self-deception involved in choosing a policy of evil (as we are doing now).

According to Sartre, self-deception is based upon (a biased view of) the available *evidence*. If we are to deploy this account, there must be something which corresponds to evidence in the case of the will attempting to make a choice of a meta-maxim based on a conception of itself as freedom. I take it that the two conceptions of freedom: autonomy (the true conception, which demands morality) and license (the false conception, which recommends evil) must constitute the relevant evidence. It might be objected that the notion of evidence is something belonging to the empirical world and yet I am claiming that the will can make use of it. In response, I would say that we must again acknowledge that we are attempting to deploy a theory of *human* self-deception to the abstract notion of the Kantian rational will (the person considered only in his free and rational aspect) and that if it is true that the will must be self-deceived to be evil and that human self-deception is based on (the abuse of) evidence and that the will is not the sort of thing that responds to evidence in the standard sense, then the self-deceiving will must be responding to some *analogue* of evidence.[66] I submit that this analogue consists simply in the two ways the will can conceive of itself as freedom (mentioned above) and upon which it relies to form its conscious self-conception as freedom.

---

[66] Perhaps talk in terms of *grounds* is less objectionable than *evidence*: we might say that the will finds grounds for taking its freedom to consist in autonomy in the true representation of its freedom as autonomy and that it finds grounds for taking its freedom to consist in license in the false representation of its freedom as license.

Sartre's account is well-suited to illuminate how the Kantian will may exploit the evidence concerning its conception of itself as freedom. As we saw in the exposition of Sartre's account, the agent's bad faith is based on the exploitation of the underdetermination of belief by evidence. This strategy works in different ways depending on the type of bad faith involved. From amongst the various types of bad faith we have examined, perhaps the first of the two types of 'bad faith in the narrow sense' in which the agent denies properties which they posses by pretending to have different ones provides the best analogy of self-deception in the will. In this type, the strategy of exploiting underdetermination of belief by evidence involves making use of similarities between what the evidence actually shows and what the agent wants it to show just as the young coquette did in taking the man's pleasant words as acts of kindness and respect rather than as part of a seduction ploy. In the same way that this woman can exploit the similarity between kind words and seductive ones, the will can exploit the similarity between freedom correctly represented as acting on one's own law (as one does in autonomy) and freedom misrepresented as acting on one's own behalf (as one does in license) in order to misconstrue it in this latter way.

However, both the evidence used and the approach to the evidence in this case ought to be regarded as unpersuasive. We saw that the way Sartre avoided falling foul of the Deceiver Paradox here was through the claim that bad faith is a project and projects alter the way we see everything: the project of bad faith issues in the agent's regarding both the unpersuasive evidence and his disingenuous handling of it as acceptable. As we saw in Webber's analysis, this is because being in bad faith is like being in a dream in which reality as such cannot penetrate and in which even one's own actions can convincingly seem other than they are. Now, we face an analogous problem in that the Kantian will must endorse a maxim of self-deception (e.g. of employing a sub-standard treatment of the evidence) without seeing it as such since this would ruin the project of self-deception (not least because this is a supremely evil project of self-deception).[67] I do not wish to claim that the will has a capacity for make-believe or fiction or fantasy which a human being in Sartrean bad faith seems to be exercising. However, there is something analogous to this which is available to Kant and which can serve to mask the maxim of self-deception and avoid the Deceiver Paradox.

This maxim is adopted by a will which (in terms of the augmented Korsgaardian conceit) has adopted the evil meta-maxim 'at the same time' (since, as I mentioned above, evil and self-deception are thought to be equiprimordial). In the context of being evil, it will seem acceptable for it to protect that which allows it to be a 'free causality' as it sees it, namely its

[67] The reader may recall from Chapter 1, that the evil status of a maxim of self-deception was what made such a policy seem impossible without a further such policy leading to a regress according to Lawrence Pasternack (1999, p.93). We will now see how this difficulty is overcome with the deployment of Sartrean bad faith to the practical philosophy.

current self-conception, thus avoiding the Deceiver Paradox. The evil will is dimly aware of itself as autonomous but it will interpret this as a threat to its current self-conception and thereby to its ability to will with almost no constraint - i.e. with *freedom* as it understands it. A maxim which fends off the threat from autonomy (which can be characterised as an unwarranted constraint on its 'freedom') is therefore acceptable in the context of evil. It can will a maxim through which it accepts the distorted view of freedom. The false conception of freedom ought to be taken as unconvincing evidence and focusing on it, an unconvincing method but the will does neither of these things because its affirmed desire to remain - as it sees it - an entirely free causality overrides any doubts it has in this regard. Thus the will convinces itself of its false conception of freedom. This avowal distracts it from its knowledge that its freedom really lies in autonomy. Since this does not involve two fully-fledged contradictory beliefs, the account avoids the Belief Paradox.

The main topic of this chapter is the adoption of the evil meta-maxim but I would like to digress from this briefly to examine self-deception regarding freedom and *particular*, evil maxims of action. Prior to this chapter, I have denied maxim-rigorism, which means I must accept that a will which, for example, has the moral meta-maxim is capable of making exceptions to this overarching policy of good and occasionally choosing particular evil actions which involves adopting particular evil maxims.[68] And since I have argued, in the present chapter, that self-deception with regard to freedom is necessary for the incorporation of license at the level of the meta-maxim, it seems correct to say that it is also required for the incorporation of it at the level of particular, evil maxims. In the case of the will which has the moral meta-maxim and carries out an exceptional evil action, the position must be that even though it must be conscious of the true conception of freedom more globally, it must also be self-deceived with regard to freedom in relation to its particular evil maxim more locally. And just as evil at the meta-maximal level masks the workings of self-deception at that level, evil at the level of a particular maxim must mask its operation at this lower level. Arguably, the will could have both a clear conception of true freedom globally and a false one more locally if during that evil activity, the licentious conception fills the will's consciousness, blocking, as it were, its more global (true) conception of freedom. In the context of that particular, evil activity, it will be made to seem as though licentiousness is freedom *simpliciter*. Arguably, since such a policy is made to seem acceptable in this way, it facilitates the agent's explaining away the pangs of conscience which mark every instance of wrong-doing.

---

[68] We need not take these maxims to be long-term policies just because they are maxims - the doctrine of the hierarchy suggests that not every maxim is a *Lebensregel* or life rule - rather, I take them to represent temporary aberrations from a commitment to morality.

4. The evil *Denkungsart*

Let us return to the main issue of the adoption of the evil meta-maxim. So far, we have discussed the application of a self-deception account to the choice of the evil meta-maxim in terms of an augmented Korsgaardian conceit but it is possible to incorporate Sartrean bad faith into the picture of the choice of the meta-maxim outlined in Chapter 2. There it was said that although ordinary maxims of action might be chosen at specific times, the choice of a meta-maxim is not. Instead, it should be conceived as made across a period of one's life. This is because it represents a commitment (to either good or evil) and a commitment cannot be constituted in one instant.[69] To combine this conception of choice with the notion of an equiprimordial self-deception regarding freedom, we must say that if it turns out that the agent is committed to evil across a period of time, then it also turns out that across that time he must be committed to taking license as freedom and to doing all those things which such self-deception requires (such as denying that his approach to the evidence is sub-standard).[70] If he fails to commit to evil, it means he did not really take license as freedom (and *vice versa*). The notion of the choice of a meta-maxim as commitment comes from Kant's discussion of the revolution to good in the *Religion* (R 6:48), so I reserve a fuller discussion of the former for when we examine the latter in Chapter 6.

In the meantime, the important points to note are as follows: the meta-maxim as the basis of an agent's model of deliberative rationality is thought to constitute his *Denkungsart* or way of thinking - the moral attitude, as it were, which he brings to individual choices. However, the proposal of this chapter is that this attitude needs to be thickened so that, in addition to its being an evil approach to choice, it has interwoven into it the elements of self-deception just mentioned. The evil *Denkungsart* is a misguided attitude about what to value, which itself contains *as an ineliminable part*, a wilful refusal to undo this attitude. It is the precise opposite of a moral attitude to choice, which, as autonomy, seeks to keep itself *free* from a self-induced reverie which would enslave it to the constraints of empirical desire.

---

[69] Since, as we will see in Chapter 6, Kant believes a commitment to morality is constituted by moral progress.

[70] The full explication of the elements of self-deception is as follows: the affirmed desire to accept the false conception of freedom, the instrumental desire to focus on that false conception (i.e., the desire to enact process of self-deception), the focus upon and acceptance of the false conception and the denial that this is a sub-standard approach to evidence.

Chapter 5
Self-conceit

In the preceding chapter, we saw that the evil *Denkungsart* comprises much more than simply the possession of the evil meta-maxim since it must also combine within it all of the elements of self-deception with regard to freedom: the affirmed desire to accept the false conception of freedom, the instrumental desire to focus on that conception (i.e., the desire for the process of self-deception), the focus upon the false conception and the denial that this is a sub-standard approach to evidence. However, there may be a further element to the evil *Denkungsart* in addition to those which I have just mentioned and it is important to understand this if we are to see what evil the agent must do to overcome it and begin on the path of moral development. There is associated, at least with the evil will - and, it seems, all wills according to Kant - a notion he calls *self-conceit*. There are unclarities with regard to this concept in the corpus, two of which have bearing on the present project. These are firstly, whether and how self-conceit is to be distinguished from self-love and (*if* self-love is a source of wrong-doing) whether self-conceit constitutes an additional perhaps more virulent source of wrong-doing. Another (related) issue is whether self-conceit is part of the sensibility of the subject or part of (the orientation of) his will. The resolutions of these problems will determine whether self-conceit is an integral part of the evil *Denkungsart* and if so, what exactly it adds to it. There is no single decisive statement of self-conceit in the corpus. However, Kant's most extensive treatment of it appears in Chapter III of the Analytic of the second *Critique* entitled 'On the incentives of pure practical reason'.[71] I will refer primarily to this.

1. Self-conceit in the second *Critique*

At the beginning of Chapter III, Kant is concerned to emphasize that 'What is essential to any moral worth of actions is *that the moral law determine the will immediately*.' (KpV 5:71) And he warns against the risky practice of allowing empirical incentives to co-operate with the moral law. This leads into a discussion of how the moral law motivates the agent by countering (two forms of) regard for oneself (*Selbstucht*). Self-regard in general is characterized in the following way: 'All the inclinations together (which can be brought into a tolerable system and the satisfaction of which is then called one's happiness) constitute regard for oneself (*solipsismus*).' (KpV 5:73) The first of the two forms of this is said to be

[71] I will refer to this simply as 'Chapter III' from now on.

self-love (*Eigenliebe, Philautia*) or 'the self-regard of *love for oneself*', also described here as 'a predominant *benevolence* toward oneself'. The second form is self-conceit (*Eigendünkel, Arrogantia*), a self-regard 'of *satisfaction with oneself*' (KpV 5:73). Kant then immediately adds further distinguishing marks to these two by claiming that 'Pure practical reason merely *infringes upon* self-love' and 'restricts it . . . to the condition of agreement with this [moral] law, and then it is called *rational self-love*.' (KpV 5:73) In contrast, it strikes down self-conceit.' (KpV 5:73) The reason given here is that,

> all claims to esteem for oneself that precede accord with the moral law are null and quite unwarranted because certainty of a disposition in accord with this law is the first condition of any worth of a person . . . and any presumption prior to this is false and opposed to the law (KpV 5:73)

and unfortunately self-conceit is or involves a claim to esteem which 'rests only on sensibility' and which 'belongs with the inclinations' (KpV 5:73). Shortly afterwards, Kant provides another contrast, saying that,

> This propensity to make oneself as having subjective determining grounds of choice into the objective determining ground of the will in general can be called *self-love*; and if self-love makes itself lawgiving and the unconditional practical principle, it can be called *self-conceit*. (KpV 5:74)

Soon after this, Kant explains in more detail how the moral law's humiliating self-conceit is associated with a (painful) feeling of respect and how when I am confronted with the example of 'a humble common man' (KpV 5:76-77), whose simple, dignified uprightness strikes down my self-conceit, my spirit bows, even if I do not. One final notable feature of self-conceit, as presented in Chapter III, is that it seems to be at least part-constituted by (and perhaps even identified with) the specifically pseudo-*moral* arrogance of the 'enthusiasts' who take themselves to be 'moral volunteers'. This seems to involve taking oneself to be so naturally and effortlessly good that one need not consult the moral law to be good. Their actions, grounded as they are in inclination (such as sympathy), follow the *letter* but not the *spirit* of the law as Kant puts it here, which (in the idiom of the *Groundwork*) is to act merely *in accordance with* and not *from duty*. As we begin to unpack what Kant takes self-conceit to consist in and perhaps what he is rationally constrained to take it to be, we may begin to see whether or not moral enthusiasm is a part of it.

2. Reath on respect and his distinction between self-love and self-conceit

Andrews Reath's essay entitled 'Kant's Theory of Moral Sensibility', in pursuing its main goal of elucidating the notion of respect for the law, also sheds some considerable light on how we ought to understand the notion of self-conceit, at least as it appears in Chapter III. Reath's key characterization of respect is what he calls the 'intellectual' or 'practical' aspect of respect. He draws this from a footnote in the *Groundwork*, in which Kant says 'Immediate determination of the will by means of the law and consciousness of this is called respect' (G 4:401n.). This is the familiar notion of respect for persons which we owe them as rational beings capable of setting their own ends. In styling this aspect as 'practical', Reath wishes to emphasize that to act from it is intellectually 'to recognize the moral law as a source of value, or reasons for action, that are unconditionally valid and overriding relative to other kinds of reasons' (2006, p.10). The practical aspect of respect is to be contrasted with another characterization (also from G 4:401n.): the *feeling* of respect: a 'feeling self-wrought by means of a rational concept'. He says that both these characterizations reappear in the second *Critique*.

Firstly, Reath explicates the negative way in which respect is thought to be an incentive to duty. The thought is that the moral law, in infringing upon those inclinations opposed to duty is, thereby, an incentive to duty since, 'whatever diminishes the hindrances to an activity is a furthering of this activity itself' (KpV 5:79). He then goes into more detail about how respect operates by clarifying the two characterizations of it and their relation to one another, something which Kant struggled to do in Chapter III. He argues that although Kant seems to want to identify them in Chapter III, they must be thought of as separate. Firstly, the feeling of respect only arises once counter-moral inclination has been checked and so the feeling presupposes this checking. Secondly, Reath points out that in discussing the moral incentive in this passage, Kant emphasizes the point that there is 'no antecedent feeling in the subject that would be attuned to morality' (KpV 5:76) thus distancing the practical philosophy from moral sense theory. Because of this concern, Kant's theory cannot have it that we are guided to moral acts by a feeling and for Kant there cannot be a moral incentive which does not involve an *intellectual* recognition of the law. Reath thinks these considerations explain certain obscure remarks Kant makes in this same passage:

> respect for the law is not the incentive to morality; instead it is morality itself subjectively considered as an incentive inasmuch as pure practical reason, by rejecting all the claims of self-love in opposition with its own, supplies authority to the law, which now alone has influence (KpV 5:76)

Presumably, Reath is thinking that Kant means the *feeling* of respect when he denies that respect is the moral incentive here and that 'morality itself subjectively considered' is the practical aspect of respect which as something intellectual is qualified to be a moral incentive. He goes on to claim that since the checking of counter-moral inclination by the practical aspect and the painful feeling arising from this would coincide in the agent, it is tempting (though wrong) to think that they are the same. The moral law opposes contra-moral inclination not by force through a feeling but through an intellectual, practical representation. However, to fully understand how this opposition operates requires an examination of the two forms of self-regard.

Drawing on characterizations of *love* and *respect* for others given in the *Doctrine of Virtue*, Reath attempts an initial sketch of what self-love and self-conceit consist in. He argues that since love for others concerns the satisfaction of their ends and their well-being, self-love (unsurprisingly) is a concern for one's own welfare and the satisfaction of one's own desires. Similarly since respect addresses itself to 'worth, esteem, dignity, or how a person is regarded by others' (Reath, 2006, p.15), self-conceit would then be 'a desire to be highly regarded, or a tendency to esteem oneself over others.' (ibid.). Although he does not quote it, Reath seems to be thinking of a passage in the *Doctrine of Virtue* in which Kant distinguishes self-love and self-conceit in the following way:

> *Moderation* in one's demands generally, that is, willing the restriction of one's self-love in view of the self-love of others, is called *modesty*. Lack of *such moderation* (lack of modesty) as regards one's worthiness to be *loved* is called *egotism* (*philautia*). But lack of modesty in one's claims to be **respected** by others is *self-conceit* (*arrogantia*). The *respect* that I have for others or that another can require from me . . . is therefore recognition of a *dignity* (*dignitas*) in other human beings, that is, of a worth that has no price (MS 6:462).

He then draws on Kant's *second* distinction between self-love and self-conceit from Chapter III (KpV 5:74)[72] and reaches the rather startling conclusion that there appear to be *two* sources of wrong-doing rather than the single one which I, in this study, have been taking there to be (i.e., the prioritization of self-love over duty). Reath on the other hand argues that self-love and self-conceit are distinct sorts of contra-moral attitude. Self-love is a form of 'general egoism': since self-love is said to be the 'propensity to make oneself as having subjective determining grounds of choice into the objective determining ground of the will in general' (KpV 5:74), it is thereby 'a susceptibility to treat one's inclinations as objectively good reasons for one's actions, which are sufficient to justify them to others.' (Reath, 2006, p.15) Since I am able to see that everyone else has the same susceptibility, I do not consider

[72] Recall that the *first* distinction in that chapter was that on KpV 5:73: there, self-love is said to be 'a predominant *benevolence* toward oneself', and self-conceit is a self-regard of 'satisfaction with oneself.'

their selfish pursuit of their own desires as grounds for complaint. I love myself more than anyone else and I take this as a reason to prefer the satisfaction of my desires to that of anyone else's but I recognize that everyone else has the same attitude. It is rather like the cynical footballer who takes himself to be justified in doing anything it takes to win[73] and in recognizing this attitude in others, *qua* cynical player, *inwardly* does not protest about the fact that everyone else does the same.[74]

As regards Reath's analysis of self-conceit in this passage (KpV 5:74), he draws on Kant's notion that it makes self-love 'lawgiving' to argue that it 'goes a step further [than self-love]' (Reath, 2006, p.15). In contrast to the general egoism of self-love, it is a 'first person egoism, in which I act as though *my* inclinations could provide laws for the conduct of others: it expresses a desire that they serve or defer to my interests.' (ibid.) Reath modified this view in an appendix in the revised version of the paper (2006) claiming that the conceited agent would merely expect others to defer to her interests rather than also actively seek to serve them (ibid., p.24).

Although already reasonably clear, Reath later explicates the sort of esteem or respect demanded by the conceited individual, thereby shedding further light on what self-conceit is (ibid., p.17). Firstly, he reminds us that the this person makes claims to esteem which precede accord with the moral law and are therefore the sort of demands which are never warranted. The question then is what would be an acceptable, morally grounded demand for respect. This would not be for honorific respect: that which might be accorded to one who has done something especially significant and good, since whilst others may pay it to you, it is not the sort of respect you can demand. The only sort of respect owed to you (and to all people), and which correlatively you can demand of others, is respect for you as a rational and moral[75] being. Since everyone is such a being, it is an inherent feature of this sort of respect that it is owed to everyone equally and that consequently, if it is paid, then each person - and, as a result, his ends - are respected equally. Self-conceit, on the other hand, 'is a claim to deserve priority (resembling a demand for honorific respect) that implicitly treats your inclinations as special sources of reasons or value. It seeks a form of personal worth attainable only at the expense of others' (ibid.).

---

[73] For example, cynical players will sometimes commit a foul which denies the opposition a clear goal-scoring opportunity even though they know such a foul is a sending-off offense because the punishment does not outweigh the benefits.

[74] *Qua cynical* player one may *outwardly* protest very loudly about cheating or gamesmanship but this is of course itself just more gamesmanship.

[75] 'Moral' in the sense of capable of morality rather than one who behaves morally.

The point of the earlier analysis of respect in terms of its practical aspect versus the feeling of respect now comes into play: self-conceit is a misguided view about what to take to have incomparable *value* (i.e., myself and my own ends) and the practical aspect of respect for the moral law strikes down this view by confronting the agent with the truth about what incomparable practical value actually consists in. I take it that Reath's point is that respect for the moral law and self-conceit share a common language (a language of value and respect) which allows the former to communicate to the latter, as it were, how it has erred. This would not be possible if the operative conception of respect for the moral law were the mere feeling, since a feeling (even one with a rational origin) is dumb and cannot communicate.

3. Morgan's account of self-conceit

3.1 Self-conceit is universal

In an unpublished draft book chapter,[76] entitled 'Kant on self-conceit', Seiriol Morgan gives an account of what he takes the place of self-conceit to be in Kant's moral psychology, making use of some of the results of Reath's article but also rejecting some of its major claims, as we shall see. One of the first points Morgan makes is that 'Kant thinks at least some measure of self-conceit is present in all human beings' (2006, p.2). His evidence for this is a statement Kant makes in Chapter III, in which he says, 'Now, what in our judgement infringes upon our self-conceit humiliates. Hence the moral law unavoidably humiliates every human being when he compares it with the sensible propensity of his nature.' (KpV 5:73) Morgan also reasons that self-conceit in the second *Critique* must be taken to be 'a basic structural feature of human agency' (and something therefore universal) rather than 'a specific idiosyncratic character trait' (2006, p.6) because it would be 'verging on the ridiculous' (ibid.) if Kant were claiming that the sources of wrong-doing are divided into selfishness and conceitedness and correlatively, that being a 'conceited ass' is a special way of being bad with a unique relation to the moral law. We would be left wondering why other vices were not also singled out as having this status. In addition, it may be that Kant takes self-conceit to be universal because it is 'something that must be postulated in order to account for some feature of the way human beings act' (ibid.). The thought that what is involved here is an empirical generalisation is dismissed since there is no indication of this in Chapter III. He thinks we therefore should suspect that Kant thinks he has some *a priori* argument for the universality claim and that 'Self-conceit will turn out to be the condition of the possibility of some phenomenon of human agency' (ibid., p.7).

---

[76] This draft chapter was placed on Seiriol Morgan's page on the website of the Department of Philosophy of the University of Bristol at http://www.bristol.ac.uk/philosophy/department/staff/sm.html in 2006. I therefore cite it as 'Morgan, 2006'.

3.2 Self-conceit is *the* source of wrong-doing

Morgan concedes that it is quite natural to understand Kant's division of self-regard into self-love and self-conceit as an indication that the latter is a particularly pernicious form of self-concern which is always bad and that the former is a form of self-concern which can be good or bad depending on what it prompts us to do and how we react to this. According to this view, self-love merely suggests I pursue my own happiness. When it prompts me to do something which does not violate the moral law, I will obviously not be doing anything wrong in choosing to do this. Conversely, self-love may prompt me to do something which happens to violate the law and in this case, pure practical reason checks self-love. If I nevertheless choose to act on an immoral maxim, self-love is then the source of wrong-doing.

He rejects this picture of a binary division between sources of immorality for the following reasons: he takes it that Kant has two conceptions of self-love from the *Groundwork* onwards. The first of these is the 'simple propensity to take our inclinations as providing us with candidate reasons for action' (ibid., p.9) and he cites the footnote from the *Groundwork* on (G 4:401) as an example of this use. I presume he is thinking of Kant's claim there that 'Respect is properly the representation of a worth that infringes on my self-love' (G 4:401n.) The thought is perhaps that if self-love *here* is something infringed upon - i.e., *limited* by respect for the law - it could not be something that is inherently bad since if it were, the law would not seek merely to limit it as though there were a level of badness acceptable to the law (which is preposterous). I think another reason for supposing that Kant must have a non-evil conception of self-love is that the evil meta-maxim is said to consist in the *prioritization* of 'incentives of self-love and their inclinations' (R 6:36) over the incentive of duty. But if it is this *prioritization* of self-love over duty which is evil itself, it cannot be the case that self-love *in this context* is evil itself. Morgan says that the second conception of self-love is that of the evil supreme practical *principle*: i.e., what I have been referring to in this study as 'the maxim of self-love' or 'the evil meta-maxim'. In support of this, he tells us of a reference to it in the second *Critique* (KpV 5:22) and we can find other instances of it in that work. For example, a little further on in the Analytic he explains why one 'could by no means pass off the *principle of self-love* as a *practical law*' (KpV 5:26). Elsewhere in the Analytic, Kant says, 'The direct opposite of the principle of morality is the principle of *one's own* happiness made the determining ground of the will' (KpV 5:35).[77]

The next move is to point out that when Kant compares self-love and self-conceit in Chapter III we must construe self-love in one of the two ways outlined above. Morgan argues that if we take it in the first sense, then it is not a source of immoral action or any action at all

[77] Although he calls it 'the principle of one's own happiness' here, in the same passage, one page later he uses the term 'maxim of self-love' (KpV 5:36) and it is clear from the context that these are the same.

because all it does is merely 'to incline the agent towards acting in ways that will gratify her inclinations.' (2006, p.9) As a mere incentive, it cannot be the source of actions since this source must be the *free choice* to act on an incentive. If, on the other hand, we take self-love in the second sense, then it *is* the source of immorality because it is the policy of prioritizing self-interest over duty and all intentional action is grounded on a principle in Kant. But, as Morgan points out, 'this is just how Kant defines self-conceit in the Analytic' (ibid., p.10), where Kant says, 'if self-love makes itself law-giving and the unconditional practical principle, it can be called self-conceit' (KpV 5:74). Thus, Morgan concludes that, 'the only sense of self-love in which it could be a source of action equates self-love with self-conceit, and collapses the putative binary distinction between the sources of immorality.' (2006, p.10) According to Morgan, this means we cannot interpret Chapter III as setting out two different sources of immorality. There is only one: self-conceit. This represents a shift from selfishness as the source of immorality in the *Groundwork* to an inflated sense of self-esteem as this source in the second *Critique*.

3.3 Accounting for self-love in Chapter III

Given that Morgan is claiming that self-conceit is the source of wrong-doing, however, he acknowledges that we must account for what self-love is in Chapter III if not a source of immorality. It would strengthen his position if it could be shown that in the two instances[78] in Chapter III in which self-love and self-conceit are compared, Kant is using the term 'self-love' in the first sense given above, i.e., the 'simple propensity to take our inclinations as providing us with candidate reasons for action' (Morgan, 2006, p.9),[79] rather than in the second sense of it, i.e., as the prioritization of self-love over duty, i.e., a full-blown principle through which the will directs itself[80] to choose evil maxims of action. If it were meant in this second sense it would undermine the view that self-conceit is *the* principle from which wrong-doing emanates, since Kant would hardly place two principles which reduce to one another side by side in the text and give them different names.

Let us, then, turn to self-love as presented in Chapter III. As Morgan points out, on page 74, Kant characterizes self-love as 'the *propensity* to make oneself as having subjective determining grounds of choice into the objective determining ground of the will in general' (KpV 5:74; emphasis added) and this can be taken as evidence that self-love is meant here in the first of the two senses given above, since it is here a mere propensity as opposed to a principle of the will and is therefore not in itself evil. It is our pathological nature orienting us

---

[78] The first instance of a comparison in Chapter III is on KpV 5:73 and the second on KpV 5:74.

[79] I will refer to self-love in Morgan's first sense as 'the propensity'.

[80] I will refer to self-love in Morgan's second sense as 'the principle'.

towards our happiness and this cannot be identified with a *free choice* to indulge it.[81] As we saw earlier it is only a free choice of the will that could be evil. In contrast to this, Kant *does* say of self-conceit in this passage that it is an 'unconditional practical *principle*' (KpV 5:74; emphasis added). As a principle, self-conceit is something freely chosen, thereby fulfilling a necessary condition of its being a source of immorality.

The other comparison of the two forms of self-regard appears one page earlier (KpV 5:73). Here, the contrasting ways in which respect relates to self-love versus self-conceit also constitutes evidence that self-love is not meant to be taken as a source of wrong-doing. Recall that respect *strikes down* self-conceit whereas (just as in the footnote (G 4:401n.) from the *Groundwork* cited earlier), it merely *infringes upon* self-love. Now a policy of evil (or an attitude associated with it) is just what one would expect to be struck down by respect for the law, whereas (as we saw earlier) something which it infringes upon cannot itself be evil because it would be ridiculous to suggest that the law seeks merely to limit evil as though there existed a level of evil acceptable to the moral law. This means that the conception of self-love with which Kant is working in Chapter III cannot be that which is equivalent to evil and the source of all wrong-doing, which leaves it open for self-conceit to be just this.

I believe that these considerations explain why, at one point, Reath gets rather muddled when trying to explain self-love. There is a paragraph in which he unwittingly seems to be trying to take self-love in both of the senses outlined above. Reath begins the paragraph by saying that it is 'a concern for well-being' and that 'in recognizing no moral restrictions, self-love makes the moral law a subordinate principle.' (Reath, 2006, p.16) He says that, in the idiom of the *Religion*, it reverses 'the moral ordering of incentives'. (ibid.) Thus far, it is clear that this is self-love in the second of the two senses discussed above - that of the evil supreme maxim. But Reath immediately goes on to say that 'It follows that what is bad about self-love can be corrected when restricted by moral concerns.' (ibid.) And 'When it does so, self-love can become good.' (ibid.) But these last two comments are only true of self-love in the first sense - understood as the, 'innocent' propensity, as it were, to take our inclinations as candidate reasons. Since Reath fails to distinguish between the two senses of 'self-love', he makes a claim about self-love-as-evil-principle which can only be made about self-love-as-innocent-incentive: he ends up inadvertently saying that through restraint by morality, *evil itself* can be turned into good! The only way to avoid this absurd conclusion is to take self-love in Chapter III to be the innocent incentive of self-love, which, when limited by respect

---

[81] By 'pathological nature', I take it that Morgan means some aspect or side of the will which favours the satisfaction of empirical desires rather than meaning that part of the whole person - sensibility - which consists of the inclinations. I take it this way because he later (2006, p.15) characterizes 'the pathological self' as something which can have a *point of view* (with regard to the choice to satisfy inclinations) and which can *strive* to make its *claims* acknowledged. These are things only practical reason - the will - can do as opposed to unthinking sensibility.

for the law is a part of goodness - i.e., part of the moral meta-maxim constituted by the prioritization of duty over *self-love*.

3.4 Self-conceit is not a quality of sensibility

Morgan reasons that if self-conceit is the incentive[82] which underlies wrong-doing, then it must be universal to human beings because no one possesses a holy will and everyone is therefore capable of wrong-doing. But there remains the task of showing why *this* is the evil incentive. If it is the evil incentive, it cannot be as Kant (inexplicably) maintains in Chapter III, a quality of sensibility. As Morgan points out, it is Reath's view that it is part of sensibility. We can see why Reath thinks this since it is the first significant thing in Chapter III which Kant says about self-regard in general (of which self-conceit is a form). Recall that self-regard is said to be, 'All the inclinations together (which can be brought into a tolerable system and the satisfaction of which is then called one's happiness)' (KpV 5:73). But Reath's view is not limited to this sensible conception of self-conceit: he struggles admirably to keep faith with Kant's other claim, irreconcilable with this one, that self-conceit is also something having to do with (or perhaps partly being) a set of quasi-intellectual capacities such as being a form of self-*regard*, i.e., as something that can 'manifest itself as *interest* in one's own welfare' (Reath, 2006, p.15; emphasis added). It is clear (to me at least) that it is quite reasonable not to be pulled in two directions, as Reath is, and to abandon one of these conceptions. Since Kant takes self-conceit to be law-giving and wants its most basic function to be as a form of *regard*, i.e., something which belongs to intelligence and not sensibility, I think we must come down on the side of its being part of the former and not the latter though it may refer to the latter, i.e., it can be a regard *about* sensibility.

Morgan also reaches the conclusion that self-conceit cannot be 'an inclination or other determination of sensibility' (2006, p.12) He asks us to consider whether self-love - the other form of self-regard - could be an inclination. To answer this, he reminds us that self-love is 'the propensity to make one's subjective determining grounds of choice into objective determining grounds of the will, that is, to act to satisfy one's desires,' (ibid.) If this propensity were itself a desire, then in order to give this desire some role in the theory, the thought would have to be that ordinary desires cannot motivate on their own without our having some extra desire that ordinary desires be satisfied. But this meta-desire (*qua* desire on this picture) would itself need a desire 'above' it to be operative and this requirement

---

[82] Admittedly up until this point Morgan has (quite reasonably) taken self-conceit to be a principle (e.g., p.10) - i.e. self-love made law-giving - but now (p.11) without explanation he calls it an incentive. However, it is entirely appropriate to think of it as both of these things since that which is an incentive can be endorsed by the will to produce a corresponding principle.

gives rise to a regress. Of course, according to the Incorporation Thesis, it is true that desires do not motivate on their own. But this is because they must be *endorsed by the will* not because the agent needs some empirical meta-desire that ordinary desires be satisfied. In addition, we can find textual support for Morgan's claims from the *Religion* that self-love is not a determination of sensibility: when outlining the three predispositions (which we met in Chapter 3), Kant says, 'The predisposition to humanity can be brought under the general title of a self-love which is physical and yet *involves comparison* (*for which reason is required*)' (R 6:27; emphasis added). Also, referring back to this (second) predisposition one page later he says, 'the *second* is rooted in . . . reason' (R 6:28).

Similarly, Morgan thinks that if *self-conceit* were a quality of sensibility, the thought would presumably be that 'self-conceit is the collective name for a category of desires with counter-moral content. Desires that others should suffer . . . desires to commit . . . rape, murder or theft' (2006, p.12) and so on. This might indeed be the way in which one who takes self-conceit to be a quality of sensibility cashes it out but alternatively they might suggest that it is a single, overarching empirical desire to be accorded excessive respect, i.e., a desire that others pander to all of one's 'ordinary' desires no matter whether and how seriously they violate the rights of others. Either way, Morgan is right that self-conceit-as-desire can never cohere with Kant's conception of rational agency. The problem is that self-conceit is meant to be legislating but on either of the two readings just given, it is mere desire and 'no desire is itself legislating' (ibid.). Once again if desire determined us directly, we would not be responsible for our actions and so 'something beyond the inclinations themselves is needed to account for the influence of sensibility upon us in a way which preserves our responsibility for acting on it.' (Morgan, 2006, pp.12-13) The source of self-conceit is not in sensibility but nor does it lie in 'the agent's rational powers considered in themselves', Morgan thinks, because 'practical reason operating correctly simply informs the agent that he has overriding reason to act according to the categorical imperative' and Kant believes everyone knows the law (2006, p.13). Nor can self-conceit be blamed on a breakdown in the agent's rational capacities since Kant seems to rule this out and, in any case, this would render the subject not responsible. By process of elimination, Morgan concludes that self-conceit must be located in the will.

4. License and self-conceit

Self-conceit can be understood as the policy of respecting ourselves excessively and correlatively failing to accord others the respect we owe them - i.e., what amounts to a policy of prioritizing our choices over those of others. As we have just seen, self-conceit is located in the will. Morgan's next task, then, is to discover what incentive we could have to adopt this policy and how this could be an incentive of the will - i.e., one which emanates

from within the will itself. Eschewing the details,[83] his strategy is to argue that there can and must be such an incentive of the will and that it must consist in the misrepresentation of freedom as license: since there are immoral acts and these must be freely chosen, the will must offer itself an incentive to adopt immoral maxims. The source of normativity for the will is *freedom*, so this incentive must be one masquerading as freedom. In short, the incentive to wrong-doing is the incentive to license with which we are already familiar. The thought that the root of all wrong-doing is misconstrued freedom is said to dovetail with the thought of it as excessive respect for oneself since self-conceit is also said to be law-giving and both license and self-conceit entail the prioritization of one's own choices (i.e. one's own outer freedom) over the choices and the freedom of other wills. Morgan raises the possibility that we might be able to identify self-conceit with an endorsed[84] incentive to license.

However, he rejects the notion of an identity because of the worry that this conception of self-conceit as license is, as it were, not sufficiently conceited. A conceited individual regards himself as better or more important than 'at least the general run of people that he encounters' (ibid., p.19). He thinks that the trouble is that there does not seem to be anything about licentiousness itself which requires that the agent take himself to be more important than those whose competing desires he tries to thwart. He says all that is required is that he prioritize his desires over those of others and that we can conceive of someone doing this without taking himself to be better. I think perhaps one could offer Reath's general egoist as an example of this.

However, I think there is an argument available to show that license *does* require a conceited attitude and that Reath's analysis of the respect for others afforded by the good person versus that afforded by the bad one provides a clue as why this relation might exist. Let us consider an action which affects other people and their ends. When this action has moral worth, it is necessarily based on a maxim which expresses my autonomy: my true and complete freedom. In addition, since, *ex hypothesi*, this worthy action affects others, I necessarily afford them and their legitimate ends the respect which the moral law demands through that action. This is the respect I owe them as persons (as beings in whom the moral law inheres). What is more, it must be the case that the reason I respect them is *because* I am conscious that I ought to respect them. It is not possible to do something with moral worth if it is not motivated by respect for the law or, what is the same, respect for persons. In short, through my worthy act, I both non-accidentally express my freedom and non-

---

[83] This is the argument for the incentive to license which is reproduced in Chapter 3 of this study and discussed in it thereafter.

[84] He argues that self-conceit cannot be identified with the incentive to license because the former is law-giving and capable of determining action whereas the latter, *qua* incentive, cannot determine action. It would have to be endorsed (thus giving rise to a maxim) to do this.

accidentally show respect for persons. Conversely, when my action is evil, it is necessarily based on a maxim which expresses *a* licentious conception of freedom (since it has been shown that license is a condition of the possibility of wrong-doing). In addition, since, *ex hypothesi*, this evil act affects persons, I necessarily at least fail to show the respect I owe them as persons.

However, given that the inescapable moral law *commands* me to respect persons as ends in themselves, to do the opposite (to deny appropriate respect to them), arguably, always requires that I *actively disrespect* them. This is because, to do them wrong, the overriding call to respect them must be overcome, by a voice which makes a false claim to having a superior authority to that of the law. When the thought, 'They must be respected' inevitably occurs to me, arguably, it can only be countered by another thought along the lines of 'I am superior to them'. The call to respect them cannot merely be ignored without being replaced by something really opposed to it. Given the unavoidable call of duty, I cannot just fail to think about their worth, I must actively despise them, hence I must be conceited to do evil to them. We have, then, parallelisms of both freedom and respect: a moral act affecting others expresses autonomy and is grounded on respect for persons and the corresponding immoral act expresses licentious outer freedom and requires a conceited contempt for others or rather, for the moral law which inheres in them.

When considering this issue of active disrespect, it may be relevant to recall Kant's point in the *Doctrine of Virtue* that, 'Virtue = +a is opposed to *negative lack of virtue* (moral weakness = 0) as its logical opposite (*contradictorie oppositum*); but it is opposed to vice =. . .[-a][85]. . . as its *real opposite* (*contrarie s. realiter oppositum*)' (MS 6:384). Perhaps vice is a real opposite to virtue because it has to, as it were, actively overcome the ever-present call to be virtuous in the will of a rational being in order for him to succeed in being vicious. If so, I take it that this is reflected in a similar requirement actively to overcome the call to respect the moral law which inheres in rational beings if we are to wrong them.

One interesting outcome of the thought that wrong-doing requires active disrespect for persons is that it seems to preclude Reath's notion of the general egoism of self-love as a source of wrong-doing which he drew from the *Doctrine of Virtue*. As we saw earlier, according to general egoism, an agent prioritizes his desires over those of others because he loves himself more than anyone else (and understands that others see things the same selfish way). We saw that this is something Kant does seem to claim in the *Doctrine of Virtue* where 'egotism' (here also called *philautia*) is characterized as a failure to will the restriction

---

[85] The Cambridge Edition of *The Metaphysics of Morals* contains the misprint 'vice = +a', which I have corrected in the quotation.

of one's self-love, i.e., a failure to moderate one's demands but it is distinguished in this passage from self-conceit (MS 6:462). But if the argument that wronging others requires my contemning them is correct, then the mere fact of my excessive self-love will not be sufficient for my wronging them where our desires clash. It seems that if Kant wants his agents all to know the law and know that others must be respected, then he cannot have a source of wrong-doing which does not involve self-conceit. Therefore, if *egotism* in the *Doctrine of Virtue* is indeed meant to be a source of wrong-doing, it must be dropped as one.

However, the argument that license requires self-conceit does not explain why the licentious agent is conceited[86] - i.e., what reason he thinks he has to think that he is better than most people. If we cannot suggest such a reason for him, he may be debarred from a policy of license. Fortunately, Morgan suggests a relation between license and self-conceit which does supply the licentious agent's reason for being conceited.[87] He thinks that Reath's account, although flawed, may help to provide this. He argues that whilst Reath's conception of self-conceit does capture an idea of conceitedness, it goes so far in this direction as to be implausible. Recall that for Reath, self-conceit is a 'first person egoism, in which I act as though *my* inclinations could provide laws for the conduct of *others*' (Reath, 2006, p.15). It expresses a desire that they defer to my desires. Morgan's criticism of this is that this person is just not realistic: she would be constantly wondering why others were not willingly - perhaps even enthusiastically - deferring to her needs. No one (apart from one who is exceptionally deluded) is like this.

He also explores the possibility that if we instead stress what Reath says about the conceited person's acting 'as though' her desires provide reasons for others to defer to her, we could come up with a more acceptable reading of his conception of self-conceit. The idea would be that the conceited person takes it that some other-worldly arbiter has judged her and her desires to be more important than anyone else's. Whilst this does not require that the conceited individual expect others to defer to her, it does mean that we as theorists will have failed to distinguish between the general egoism of Reath's self-love and the first-personal egoism of Reath's self-conceit since both have the agent acting *as though* her desires provide reasons for others to defer to her.

---

[86] It is not possible to appeal here to Kant's doctrine that he who wills the end (license, in this case), wills the indispensible means to it (self-conceit) because even if we accepted that the agent saw that he had to be conceited in order to be licentious, this in itself would still obviously not give him a reason to think he was better than other people.

[87] As we saw above, Morgan rejects the idea that license requires self-conceit and is thereby motivated to provide an alternative connection between the two.

However, taking his cue from Reath, Morgan offers an interpretation which is both realistic and which is distinguishable from general egoism. This picture has it that the conceited person thinks that there is an objective reason entitling him to prioritize his inclinations over those of others which they do not possess. Morgan's suggestion is that he takes his having 'the strength, or the courage, or the initiative' (2006, p.23) to adopt a policy of license - 'to break free from and brush aside the claims of others, and exert [his] own resolve in the face of them' (ibid.) - as a reason to regard himself as having a greater worth than other people.[88] He also takes it that others who have not embraced license, lack this special status and therefore have reason to curb the pursuit of their desires where these conflict with (the legitimate claims) of others - i.e., the conceited agent takes it that they have 'reason to act like dutiful Kantian agents' (ibid., p.21). However, he will accept that they are being rational when they pursue their own ends. Whilst he accepts that, from their point of view, they have no reason to defer to him (any more than morality requires), and that they are entitled to try to avoid having their desires frustrated by him, they would be wrong to think that 'someone like him ought to accept that he should not try to dominate them' (ibid., p.22).

One difficulty with this is it seems to require that the individual *revel* in the fact that he has chosen a policy of license *as such*. If so, it is not something the licentious will which I envisaged in Chapter 4 could do because there I argued that in order to adopt a policy of license, the will must *self-deceptively* take license to be freedom, thereby avoiding taking it as license as such. *That* self-deceived and licentious will could not embrace self-conceit in the way suggested: i.e., by openly acknowledging its license as such. However, we need not suppose that the will thinks of its license *as license* in order for it to take it as grounds for regarding itself as better or more important. Instead, to that will, limitless outer freedom is freedom *simpliciter* and its preparedness to embrace it, and the crushing of other wills which it invariably involves, constitutes grounds, as far as it is concerned, to place it above those others who are not prepared to do this. Thus, this conception of 'freedom' is the ground of self-conceit without being thought of as license.[89]

However, this presents us with two related problems: firstly, it might be argued of our agent who takes license to be freedom that since freedom is universal, it might be reasonable to expect her to think that freedom is universal. But then she will not think that she is anything

---

[88] If my argument that self-conceit applies to violations of duties to oneself as well as to others, then the licentious individual may be taking pride in their ability to overcome what they might take to be fussy moral scruples about how they 'purportedly' ought to treat themselves in the pursuit of empirical desires.

[89] Earlier I mentioned that Morgan moved from speaking of a principle of self-conceit to an incentive of self-conceit without explanation and that the thought must be that where there is a principle there must be an incentive to endorse it. We can now see the incentive to endorse a principle of self-conceit (i.e., the reason to do so) is the acceptance of a policy of license itself.

special in being (licentiously) free: it would not be grounds for her to be conceited. Instead, (and this is the second problem) the agent's licentious conception of freedom would only be taken by her to be grounds for an attitude of general egoism: she would think that since our 'freedom' justifies egoism and everyone is free, *everyone* is justified in being egoistic, not just her. However, just as one can possess the property of autonomy and yet fail to exercise it, the one who supposes that license is freedom will think that although everyone has this property - the potential to embrace it and act on it - not all have the audacity to do so.

If we are to take it that the conceited individual is conceited because she takes her endorsement of her conception of freedom - or as she might see it, her preparedness 'to do what it takes' - to be something noteworthy, then we might think that if she were to encounter others whom she recognised as having done the same, she must afford them the same respect she affords herself. The thought would then be that these people take themselves to belong to an elite group whose members, whilst having a contemptuous attitude towards those too weak, cowardly or unambitious to embrace freedom, as they see it, would presumably regard those in their group as equals since each recognises in all the others the same grounds for the inflated respect he affords himself. This perceived equality might suggest that they would refrain from trampling on one another's desires since it was a supposed *superiority* over others which allowed them to do this to members of the inferior group. However, whilst it might be plausible that they would respect one another more than they respect those outside the group, the idea that each conceited person would respect the desires of his peers as much as his own desires seems far-fetched. It seems more realistic to say that (their apparent grounds for equal respect notwithstanding) they would still compete with one another where desires clashed. My intuition is that bad people are not so impressed by each other's *chutzpah* that they refrain from being bad to one another. The question is whether the practical philosophy (as it is interpreted and reconstructed in this study) yields a result similar to this intuition.

It might be suggested that conceited individuals may compete with one another (despite being equals in terms of their all having embraced 'the good life' as they see it) because, being conceited, each simply does not care about what would otherwise come across as a difficult, audacious or impressive thing for the others to have done (i.e., having had the audacity to embrace this aggressive approach to life). The problem with this picture is that it precludes the mutual evaluation required to make one's status intelligible to others. No one in this group can claim the respect of the others in it because *ex hypothesi*, there is a principled reason why no one in the group is capable of affording it. Perhaps this individual is 'too' conceited for mutual intelligibility.

We are, then, left with the problem of providing an account of peer evaluation within the elite group which has the realistic outcome that any recognition on their parts of each other as members of that group would not prevent them from trying to overreach one another as we might expect them to do. Morgan has proposed[90] that in adopting this life and in taking this as something which confers a special worth on himself and others like him, the evil agent sets up a table of values which rivals that of the moral law and in terms of which he and others in his group can fair well or badly in relation to one another. Since license is the accepted conception of freedom amongst members of this group, it is the ability to pursue outer freedom, i.e., the means to acquire what one desires - including the means to crush opposing wills - which is respected. Respect is afforded according to the possession of power: money, influence over others and so on. Since power admits of degrees, one person can be superior to another and so it is easy to see how a superior may take himself to be justified in riding rough-shod over the ends of a perceived inferior even though the latter is recognised as a member of the 'elite' group. It is more difficult to explain how one who regards himself as inferior would take himself to be justified in trying to defeat his acknowledged superior. Morgan's answer is that the key to this lies in the fact that this sort of evaluation is *relative to another*. The inferior individual may recognize that he is inferior in terms of actual capability but this does not constitute a fact of the matter about who is superior. At any particular time, there is no fact of the matter in this regard since there is always something one can do to acquire more power than one's rival. Just by trying to be better, by competing, he is justified in competing since it may result in his being more powerful than the one he competes against. This idea reflects the internal logic of licentious outer freedom in that in accordance with this conception, I am engaged in a limitless outward striving and so another's very superiority to me may in itself be a spur to overcome him. In thinking this way, I may actually defeat the other and be demonstrably better than him.

Finally, in Chapter 3 we saw that the good agent who from time to time may fail to live up to his commitment to the good by carrying out an evil action, temporarily adopts an evil particular maxim, and thereby endorses license at the level of that maxim despite conceiving of freedom as autonomy at the meta-maximal level. Given the arguments presented in the present chapter regarding the necessity of self-conceit for violations of rights, it seems we must suppose that even the good agent, if he submits to a particular evil action and its associated maxim, must take on a conceited attitude in relation to those interests, at least, he must if they involve rights violations. Just as he is able to conceal the fact of the endorsement at the level of the maxim of a false conception of freedom, we must suppose that on entering into this limited conceitedness he is able mentally to block the fact that it must be avoided.

---

[90] In personal correspondence.

5. Self-conceit *simpliciter* and moral self-conceit as a mere expression of it

One last issue to be addressed is that Kant seems to frequently[91] associate an inflated sense of specifically *moral* self-esteem with self-conceit and moreover Morgan (referring to the *Moralphilosophie Collins*; VE 27:464-465) is right that he frequently seems to at least hint at an identity. We can see a clear example of this in Chapter III: when describing how the moral law humiliates, he cites the target of this humiliation as '*moral* self-esteem' (KpV 5:79; emphasis added). The problem is that it is difficult to see how this might fit in with the picture of self-conceit as a root or element of the root of wrong-doing which has been suggested. I have to agree with Morgan that it would be ridiculous to suggest that all wrong-doing must be *rooted* in an inflated sense of *moral* self-conceit in the sense that an agent who violates your rights, does so *because* she thinks she is *morally* better than you. I therefore cannot give a satisfactory account of those instances in the corpus where Kant seems to identify the two. Morgan has suggested[92] that if Kant did intend an identity, it may have been because he could not imagine anyone using anything but a quasi-*moral* self-appraisal to measure his self-worth *simpliciter*. However, we have seen and will see again soon that other criteria can be used by agents for this purpose.

It is clear (to me at least) that the practical philosophy needs a conception of self-conceit (an inflated sense of self-worth) since (I have argued that) a will must be conceited - must think itself superior to others - to do them wrong. However, we have no good reason to think that this sort of self-conceit is identical to moral self-conceit and a very good reason to deny this (as we have just seen). Let us call the former, self-conceit *simpliciter* to distinguish it from the 'moral' variety. Presently, I will argue that moral self-conceit should not be thought of as a root of evil (this root is always self-conceit *simpliciter*). At most (if there is to be a moral self-conceit at all) it is to be relegated to a mere expression of self-conceit *simpliciter*.

So far in the present chapter, the agent who is conceited *simpliciter*, has been characterized as the sort of person who aggressively pursues his desires and tramples on others' ends where these conflict with his own. However, this characterisation presupposes that he has the sort of desires which tend to bring him into conflict with others - for example, desires to be rich, or powerful and so on. But it seems perfectly possible for an individual to possess a will which is conceited *simpliciter* (i.e., take himself to be more important than most) and yet either not have these sorts of desires or at least not be preoccupied by them exclusively. Instead, there may be individuals whose wills are conceited *simpliciter* but who also happen

---

[91] In addition to the examples given Kant seems to be thinking of a specifically moral self-conceit at certain places in the *Doctrine of Virtue* (MS 6:435, 437, 460) and in the *Religion* (R 6:185n.)

[92] In personal corespondence.

to have non-aggressive inclinations such as those of compassion. Rather than exercising his supposed superiority only through aggressive competition, the conceited individual who also has some compassionate desires might sometimes pursue them by dominating others through care: i.e., by occupying the carer's superior position over those he judges to be both weaker and than him and not in competition with him.

What emerges from this is that the conceited attitude of will which is the root of wrong-doing - self-conceit *simpliciter* - should not be identified with aggressive domination since this is only one *expression* of self-conceit *simpliciter* (found in those with aggressive inclinations for, e.g., power and so on). Self-conceit *simpliciter* itself is an inflated sense of self-worth - pride based on the individual's wilful defiance of the constraints of the law and of others which may or may not be expressed aggressively. And just as aggressive domination is just one expression of self-conceit *simpliciter*, so is what might be termed moral self-conceit. This is very different from Kant's apparent conception of moral self-conceit inasmuch as Kant *seems* to suppose that the basic arrogant attitude of will (what I call self-conceit *simpliciter*) always has a (quasi-)moral character. The picture I have put forward has the clear advantage that it enables us to say that self-conceit (basic, inflated self-worth) remains a root of evil but in doing so, we are not saddled (as Kant apparently is) with the view that this must take the form of a specifically moral arrogance. However, it must take some form and we may suppose that this will very often (perhaps most frequently) be the aggressive form Morgan has dealt with.

One might reject the notion that philanthropy can ever be objectionable since even where it is not done from duty, the recipient still benefits and so, to criticise such action and such an agent seems mean-spirited and unwarranted. However, it must be remembered that to help someone out of pathological compassion and not out of practical love (the concern for others commanded by the law), gives rise to the risk of treating him as mere means to one's own pleasure (in helping them) and this is to fail to treat them with the required dignity. Just as in the case of the aggressively expressed self-conceit with which we have been concerned for most of this chapter until now, conceited philanthropy requires ignoring the call of the moral law to respect one's fellow men as equals by actively disrespecting them. Morally worthy *practical* love differs from pathological compassion in three key ways: firstly and most obviously, it is motivated by duty rather than an inclination which serves self-love; secondly, the truly moral philanthropist sees the person in need as a respected equal, a fellow rational being rather than someone to be pitied; thirdly, the good man's end is to help this respected equal to realise his potential as a rational agent rather than use him as a means to the end of his own philanthropic gratification. In addition to dominating others simply by occupying the carer's position, the morally conceited individual may also have to dominate others in a

manner more akin to the aggressive conceit where his putative wards resist his attentions. Again, in matters of care (just as in matters of money and so on), he thinks they are entitled to try to resist but would be wrong to think that a superior person such as him should not try to overreach them.

There is perhaps a middle ground between moral self-conceit which seeks domination and morally worthy practical love. This is philanthropy carried out by one who is not conceited and has the moral meta-maxim but who is nevertheless motivated by compassion. This is arguably not wrong or conceited because the intention is not to dominate (even if it may lack virtue). One difference between this person and the morally conceited philanthropist is that the former is prepared to stop if asked *because* he is asked. Another might be that he is prepared to try to examine his motives for helping.

This new conception of moral self-conceit raises certain problems regarding agents' reasons for thinking of themselves as more important than others. Earlier in the present chapter, it was supposed that the aggressively conceited individual took as his reason for his superiority his audacious willingness to pursue his desires and *frustrate their ends* where necessary. However, now that we have an alternative expression of self-conceit *simpliciter* - moral self-conceit - in which the conceited individual actually *assists* others, we can hardly claim that one whose conceit is sometimes expressed through philanthropy thinks himself better for the same reason as the aggressive individual does.

The solution lies in what is common to both the aggressive and philanthropic conceited individuals: it is their preparedness to declare themselves free to do whatever they want and to disregard the demands of the moral law or of any other will which they value. So it is the *pursuit* of license *in whatever form* which forms the basis of a table of values. In conceited philanthropy, the agent takes pride in her willingness to brush aside claims (whether or not they are asserted by the intended targets of her compassion) that she is doing anything wrong and thinks she is superior because she is prepared to respond to the compassion she feels, to overcome adversity, and 'do what needs to be done'. She perhaps may dismiss the notion that one ought to examine one's motives for helping people as high-minded, fussy or self-indulgent.

The claim, then, is that the reason for arrogantly thinking oneself superior (shared by both types or expressions of self-conceit) is one's sheer willingness to be arrogant in the pursuit of one's desires. The main advantage of this solution is that it allows agents indulging in both expressions of self-conceit *simpliciter* to find a reason for their supposed superiority since that reason is located in a feature of the *will* common to all licentious agents rather than in the

inclinations of a particular one. In addition, it can perhaps accommodate any other expression of self-conceit which emerges as well. The table of values that conceited wills use to compare themselves against others including other conceited wills involves the degree to which they are prepared to pursue licentious 'freedom': the lengths they are prepared to go to get or to do what they want. A good indication of this will simply be what they in fact have acquired or have done.

6. Conclusion

Let us clarify the position at which we have arrived. Earlier we saw that Morgan argues that self-conceit is the root of all wrong-doing. This is because (amongst other things) the principle of prioritizing self-love (in the innocent sense) over duty is the source of wrong-doing and this amounts to self-conceit as Kant defines it on page 74 of Chapter III since there, self-conceit is said to be self-love made *law-giving*. My position is that the will which accepts that freedom *simpliciter* consists in the limitless pursuit of outer freedom adopts a policy to suit this conception of freedom. That policy is the prioritization of self-love over duty. In accepting this conception of freedom and its policy, the agent equiprimordially accepts being conceited: he regards himself as better than others and is contemptuous of their claims (legitimate or not) and those of the law. If he does not accept self-conceit, license is not possible and he has not genuinely accepted the latter as freedom. It is in this way that both license and self-conceit are the root of evil and that the incentive to license is the incentive to both (as well as being an incentive to the elements of self-deception regarding freedom). It now emerges that in addition to the elements of the evil *Denkungsart* which we identified in Chapter 4: (the possession of the evil meta-maxim and the elements of self-deception with regard to freedom), the evil will also necessarily has an inflated sense of its own importance which can be expressed in various ways depending on the inclinations involved. These elements of the evil *Denkungsart* are the initial obstacles to be overcome in committing to a project of moral self-development. Whilst this is a daunting task, the fact that these elements are all related to one another may give us hope that the undermining of one is the undermining of them all.

Chapter 6

Moral self-development

In this chapter, we will examine how the agent may develop from a morally depraved individual into one who could be called virtuous. Although, as we saw in Chapter 4, according to Morgan's rational reconstruction of evil, there is nothing in the practical philosophy which constrains Kant to posit *universal* depravity, it is nevertheless conducive to the provision of a fuller account of the moral development of a Kantian agent to begin presently with an individual at this (the worst) stage of evil since this person, as it were, has it all to do. To begin to be a better person, the depraved individual must reject the evil meta-maxim (and thereby adopt the moral meta-maxim) despite the co-ordinated and concerted resistance of his combined evil *Denkungsart*. Since a vital element of this attitude is a perverse conception of freedom, it is clear that the key to taking even this first step lies in the agent's becoming conscious of his true freedom. The more difficult question of precisely how this is to be achieved is (the first) one I intend to answer. However, the challenges facing the would-be virtuous agent do not end there. We will see that the adoption of morality is not simply a matter of rejecting evil and resolving to be good. Its adoption also requires an affirmation of a *commitment* to morality consisting in moral progress. Such progress consists primarily in the pursuit of the duties of moral development (such as the pursuit of holiness). However, it seems two of these duties of moral development clash with one another, so it is important to decide whether they are compatible before we can proceed. In addition, some elements of moral development arguably require moral self-knowledge and we must establish whether this is possible given the opacity of maxims. Despite the continued threat of self-deception, it seems this is possible. Finally, we will see that the pursuit of each of the elements of moral development is either heavily reliant upon (or just is) the acquisition of virtue which again involves freedom's self-recognition. Since the self-consciousness of freedom both initiates the process and allows it to continue, it is a theme which unifies a programme of moral self-development.

1. Waking from the dream of depravity

Presently, we will see that from the very beginning, the consciousness of true freedom is central to the story of moral development. As I mentioned above, the initial task is to reject evil and adopt the moral meta-maxim, that is to say, to bring about the revolution. However, Kant raises a concern about how an evil person as such could do this good thing. In the *Religion* he asks, 'But if a human being is corrupt in the very ground of his maxims, how can

he possibly bring about the revolution by his own forces and become a good human being on his own?' (R 6:47) Kant's worry here seems to be to do with where the agent will find the motivation to bring about the moral revolution, given that he is evil. His response to this putative problem is first to assure us that because the revolution is required, it must therefore be possible (R 6:47).[93] He then claims that the evil person's focus on the predisposition to good can give him the required motivation. Contemplation of the sheer wonder of a predisposition through which we are capable of renouncing all worldly pleasure, 'though this alone can make our life desirable' (R 6:49) can inspire in us a 'feeling of the sublimity of our moral vocation' (R 6:50) and can work 'for the restoration of the original ethical order of incentives' (R 6:50).

However, whilst this feeling may be an encouragement, what is finally required is a full endorsement of autonomy as the true conception of freedom. This can be seen if we express Kant's abovementioned worry about the revolution in terms of conceptions of freedom: we have seen that the will regards as appropriate whichever policy expresses its conception of freedom since freedom is normative for it. This means a will which self-deceptively takes license to be freedom must have the evil meta-maxim and is *aware* of no reason[94] to adopt the moral meta-maxim, i.e., to will the revolution. This in turn means the acknowledgement of the will of true freedom is a *necessary condition* of the revolution. It is also sufficient for it because morality is the appropriate choice for a will that takes autonomy as freedom since it is the policy through which that freedom can be expressed.

However, the suggestion that clear consciousness of the fact that autonomy is freedom will bring about the revolution (i.e., that it is sufficient for the adoption of the moral meta-maxim) merely pushes the problem Kant raises back one step since now the question arises of how an evil will could achieve such consciousness. One possible answer is that the performance of a single moral act may furnish the evil will with a consciousness of the true nature of freedom since through that act, true freedom will be exercised. In the following passage from the second *Critique*, Kant himself also seems to be suggesting that we begin with minor good acts (or attempts at them). He says,

> this consciousness of the law as also incentive is inseparably combined with consciousness of a power *ruling over sensibility*, even if not always with effect; yet frequent engagement with it and the initially minor attempts at using it give hope of its effectiveness (KpV 5:159).

What I am proposing is that since the will is freedom itself, and since a moral act exposes it to its true freedom, an initial moral act may encourage further ones and undermine self-

---

[93] The same argument was used a few pages earlier (R 6:45) and is repeated again a few pages on (R 6:50).

[94] Although, it dimly knows that it does have such a reason.

deception with regard to freedom. If self-deception collapses under the weight of evidence that autonomy is freedom, then since freedom is normative for the will, I take it that it must adopt the moral meta-maxim, i.e., the revolution takes place. In addition, since the other elements of the evil *Denkungsart* - including the affirmed desire to take license as freedom and self-conceit - are based on possession of the evil meta-maxim, when this is overturned, they will also dissipate.

But, again, some might object that the problem Kant raises about an evil will adopting a good meta-maxim is merely pushed back yet *another* step by my suggestion that the revolution could be 'sparked' as it were by an initial *good* action which fixes the will's attention on its autonomy since the will in question is *evil*. One might think that we are faced with a difficulty similar to the one dealt with in Chapter 4, where we considered why a will would endorse the evil desire that license were freedom when this could issue in a self-deception which facilitates the adoption of the evil meta-maxim. The question now is, in short, how and why a will would go against its commitment to evil and do something good which could issue in the abandonment of that policy of evil.

The response is that the present putative difficulty is different from the one associated with the desire which brings about a choice of the evil meta-maxim in a number of important ways. Firstly, in Chapter 4, we saw that the desire to see license as freedom could potentially bring about self-deception with regard to the will's *whole conception of freedom* and is a *direct* attack on and corruption of *the root* of the power of choice. This desire is therefore not only evil but also very significant and therefore seemed to conflict starkly with a will which has overriding reason to choose morality. In contrast, the act which we are considering now, which 'sparks' the revolution is in itself small and insignificant and does not directly concern the will's whole conception of freedom nor, therefore, its orientation to good or evil. It therefore does not seem to constitute the same sort of clash (this time, with a policy of evil). True, it is still a good act carried out by an evil will but a commitment (whether to good or evil) *qua* commitment is momentarily defeasible. As we have seen the relation between meta-maxim and maxim of action is not causal - and as Allison puts it, merely 'broadly logical' (1990, p.142). In addition, if a commitment as such is defeasible, then arguably a commitment to evil is more readily contravened by an individual instance of moral willing since morality is overriding. As Kant says, in the second *Critique*, 'The maxim of self-love (prudence) merely *advises*; the law of morality *commands*.' (KpV 5:36)

Another objection concerns the interdependence of the elements of the evil *Denkungsart*. This is perhaps a two-edged sword since whilst it means that the successful undermining of one element (self-deception regarding freedom) can be the undermining of all, this very

interdependence is also a source of robustness for the *Denkungsart* as a whole. The elements *stand* as well as fall together. For example if self-deception regarding freedom comes under pressure, then the will's policy of evil and corresponding evil desire to take license as freedom may come in to play to 'shore up the defences': the evil will desires to maintain that which allows its expression of freedom as it sees it and its false conception of freedom is one of those things and will be defended. This is why I claim that any consciousness that autonomy is freedom will be fleeting and isolated at first, since it will be suppressed by the evil will. We can imagine that the evil agent who has a rare instance of consciousness of what it is to act autonomously and be free of the demands of his inclinations may find that something in him fights back so that he reminds himself that if he opts not to live according to his desires, then he is resigning himself to all of those acts which are almost always unprofitable and can be variously arduous, time-consuming, frightening, painful, boring, expensive and so on and he is of course always at risk of having to forgo all those things which make him positively happy as well as the happiness he derives purely from the freedom he now has to pursue them. To live such a life would be an unwarranted curtailment of freedom as he sees it: namely unlimited outer freedom.

Also, as we learnt in Chapter 5, his embrace of this life and this conception of freedom give him pride in himself. As far as he is concerned, it is what makes him and those like him 'a cut above' the rest. 'Ordinary' people are thought to be too lazy, or cowardly or stupid to live a 'free' and uninhibited life or to accept that one needs to 'do what it takes' to flourish and even merely to 'get by'. In fact, since the life of license is worthy of respect from the point of view of the licentious individual because it requires a certain audacity or 'strength' of will, any thoughts that true freedom lies in autonomy may be misconstrued by the evil individual as temptations to abandon a life of self-affirmation from cowardice or weakness. The one who, in reality, deceives himself into thinking freedom is license and that imposing oneself on others is to be respected may reject thoughts which could lift this self-deception, by falsely dismissing *them* as self-deceptive: as driven by cowardice and so on. This is all the more convincing since the life of license really does require audacity (rather than what Cicero would call courage)[95] since a life dominated by the violation of the rights of others is likely to be a life dominated by *their resistance* to such violations or at the very least resentment and moral outrage. In addition, if his self-conceit also sometimes takes on a quasi-moral form, he again regards himself as more important than others but, at the same time, one who deigns to care for his fellow man. He is therefore already moral in his eyes and any call to examine his motives is seen as fussy and unwarranted.

---

[95] In *On Obligations*, Cicero makes a useful distinction between that greatness of spirit which is coupled with justice and that which is not but is instead coupled with the pursuit of narrow self-interest. Only the former can be called courage. The latter is mere 'recklessness'. (1. 62-64)

There is another possible obstacle to overturning the evil *Denkungsart*: although the licentious individual *qua* will must be supposed to be detached from his desires like everyone else in Kant, it may be more difficult for him to see that he is free in this way (and thus that is he free to choose to incorporate duty into his maxims). One who has endorsed the incentive to license may come to see his empirical desires as more integral to himself *qua* rational chooser than they in fact are in the Kantian scheme of things. The thought is that he may come to see things this way because license is itself a non-empirical *desire* emanating from the *will itself*. But it is also exclusively associated with the empirical desires which the licentious will values above duty. The desire of license may make what is empirical and not at all a part of him *qua* will, seem as though it is less distinct from him *qua* will. These considerations may provide Kant's supporters with the beginnings of a response to Williams' charge (outlined in Chapter 1) that Kant's purported mistake is to overlook the 'fact' that 'practical deliberation . . . involves an *I* that must be more intimately the *I* of my desires than this [Kant's] account allows.' (1993, p.67) This was supposed to mean that any attempt to will on the basis of moral grounds (which refer to no empirical interest) may cause the diremption of the self where such willing conflicts with the agent's most cherished non-moral interests, interwoven as Williams supposes they are with the agent *qua* 'deliberative *I*. However, contra Williams, we may say that when one's desires seem part of oneself *qua* will, *qua* intelligence, it may be because those are the only interests one's *Denkungsart* allows one to see (or see clearly) and they thus seem almost to exhaust who one is.

There is no doubt that the factors opposing the undoing of self-deception are formidable but we should not think this makes it impossible. That this picture of the *Denkungsart* with its mutually supportive elements makes overturning it out to be a difficult thing to do is no cause for objection to that view of it. On the contrary, the recalcitrance of evil in this Kantian picture may be reflected in reality if it is true (as I suspect it is) that most bad people stay bad and that the fundamental reorientation of a moral outlook which is supported by self-deceptive narratives is something rare.

2. The revolution: resolving to be good, committing to morality and the adoption of the moral meta-maxim

We now turn to the revolution and an examination of the key passage from the *Religion* upon which is based the notion of the adoption of a meta-maxim understood as commitment. We will see that, according to Kant, the commitment to morality is constituted by moral progress. We will then be in a position to see (later on in this chapter) how this commitment is dependent on a consciousness of freedom. The will's conception of freedom is normative for it, so it must adopt that policy which allows the expression of that conception of freedom.

The one who truly manages to undo self-deception is conscious that *autonomy* is freedom and must adopt the moral meta-maxim;[96] he must inaugurate the revolution - the beginning of the difficult process of moral development. One might object that if an erstwhile self-deceived person is un-self-deceived from a particular time, arguably, he will become conscious of autonomy as freedom and will ineluctably adopt the moral meta-maxim (i.e., become a good person) *at that time*. However, this seems to contradict the conception of the choice of a meta-maxim I have advocated so far in this study: that it represents a commitment, which as such, cannot be realised in an instant.

The origin of this idea of a choice of meta-maxim as commitment as well as the answer to the puzzle I have just described can be found in Korsgaard's analysis of a key passage on the revolution from the *Religion*. There, Kant says,

> If by a single and unalterable decision a human being reverses the supreme ground of his maxims by which he was an evil human being (and thereby puts on a "new man"), he is to this extent, by principle and attitude of mind, a subject receptive to the good; but he is a good human being only in incessant laboring and becoming i.e. he can hope - in view of the purity of the principle which he has adopted. . . - to find himself upon the good (though narrow) path of constant *progress* from bad to better. (R 6:48)

We can see here that Kant distinguishes between the *resolution* to be good and *being* good. One qualifies as good 'only in incessant labouring and becoming' (R 6:48). I take it that it is this statement which leads Korsgaard to think that one qualifies as a good person only by realising one's commitment to morality and this is done by making moral progress.[97] However, she makes the further point that 'There is a sort of backwards determination in the construction of one's character' (1996a, p.181) so that if it turns out that one is committed (by making progress), then that makes one's resolution sincere - i.e., it turns out that one really did adopt the maxim at the point of the resolution. This means there is no discrepancy between my claim that the moral meta-maxim is adopted at the point at which one's self-deception regarding freedom is lifted, on the one hand, and the notion that the adoption of a meta-maxim is constituted through a commitment taking the form of moral progress on the other.

I would like to address two points in the passage which at first may seem at odds with my position. Firstly, I say the adoption of the meta-maxim is constituted by a commitment in the

---

[96] The will's conception of freedom is normative for it, so that conception of it (either autonomy or license) is both necessary and sufficient for the adoption of the corresponding meta-maxim (respectively, good or evil) which allows the expression of that conception of freedom.

[97] What I take to be the main elements of moral development in which this progress must be made are set out in Section 3, below.

form of progress. But in the passage, Kant says the agent reverses the supreme ground of his maxims *by the decision*. However, this is compatible with my view because Kant is thinking of a *sincere* resolution - he calls it 'unalterable' by which I understand 'firm'[98] and also it could not be a *decision* if it were insincere - and in the story I have expounded, if the resolution turns out to have been sincere, the moral meta-maxim turns out to have been adopted from the point the resolution is made.

Secondly, Kant's view seems to differ from mine on whether the person is good at the point of the resolution, saying that when he 'reverses the supreme ground of his maxims,' he is merely 'receptive to the good' (R 6:48) rather than good *simpliciter*. However, I think this is merely a reflection of the fact that one is not yet good in the sense of not yet having developed one's goodness at the starting point. We can see how when a person has just decided to be good, most or even all of his maxims of action will still be evil and self-serving (and so it is tempting to say that this person is merely receptive to the good rather than good *simpliciter*). But I can agree with this and still maintain that there is a sense in which one is good at this point: in terms of sincere intention to become a better person, the potential to progress and so on. Moreover, if Kant does not call the person who has adopted the moral meta-maxim good, then he must, by disposition-rigorism, call him evil, which is an awkward result. Nevertheless, I think Kant would not object to the thought that this person is good at the point of his resolution in the sense that he sincerely wills to do good and to improve[99] *and* that this is how it turns out.

Since I take it that one who is un-self-deceived must have the moral meta-maxim and that in having this meta-maxim, it must be the case that he is committed to morality, I must also hold that one who is un-self-deceived must *inevitably* turn out to be committed and make progress. This may seem somewhat unintuitive. However, I think it seems less so when it is pointed out that if one fails to progress, then it turns out that one did not really take autonomy as one's conception of freedom and that one's resolution was insincere. Perhaps another way to look at it is that if one truly accepts that autonomy is freedom, then one is *already the sort of person* who will make progress and thereby turn out to be committed and is therefore already good at the point of the (sincere) resolution. It will emerge in the

---

[98] I think that Kant cannot literally mean that this sincere decision is *unalterable* as this would contradict his view that a fall from good to evil is possible e.g., R 6:45. I think that he must mean that the decision is non-literally unalterable - e.g., *firm*. Drawing on the ambiguity of the scope of the word 'unalterable', Seiriol Morgan has also pointed out (in personal correspondence) that Kant means the agent resolves to be unalterably good rather than unalterably resolves to be good. This I think gives the same or very similar result: i.e., that Kant means the agent has decided to strive to be good with complete sincerity.

[99] Kant explicitly states earlier in the *Religion* that one who has the moral meta-maxim is good: 'Only when a human being has incorporated into his maxim the incentive implanted in him for the moral law, is he called a good human being' (R 6:45n).

Conclusion of this study that this can be explained in terms of freedom's determination to free itself fully. I realise that this notion may be unclear now but I hope its meaning will become clearer by the end.

3. The core elements of moral development

We have seen that the adoption of the moral meta-maxim is initiated by an awareness that autonomy is freedom and that this adoption is then fully constituted by the realisation of one's commitment to the good by making progress in a programme of moral development. This programme consists in a number of elements (mainly duties) detailed below. For the remainder of this chapter, we will be concerned with giving an account of what the principal elements of such a project should be, and (in some cases) whether they can be pursued at all and if so, how these ends might be pursued. It will emerge that all of the various elements of moral development are facilitated by virtue itself and that it part-constitutes many if not all of the intended end-states of these elements. Since we will see that virtue is acquired through an increased awareness of the extent of our capacity for *inner freedom*, it turns out that affirmation of our commitment depends on this consciousness. Consciousness of freedom is, then, central to both the initiation and the affirmation of the entire moral development project.

Arguably, the most central elements of moral development are those of the duty to uproot evil maxims (and resist new ones), the duty of moral perfection (the duty to do all of one's duties), the duty to cultivate those empirical motives which can co-operate with duty (especially compassion), the duty to be holy, the ability to reassure oneself of one's moral progress and perhaps most importantly, the duty to acquire virtue. The first task is to establish the moral point of these elements. This can be done without much difficulty in some cases. Others will require lengthy discussion. Let us deal with two of the less troublesome elements first: the duty to uproot evil maxims and the duty of moral perfection.

The requirement to uproot evil maxims arises in the following way: it is reasonable to suppose that an evil person will have some (perhaps many) evil maxims. Since one who inaugurates a sincere revolution is evil prior to it and good afterwards, and since there is no reason to suppose that his evil maxims will automatically dissolve by his merely resolving to be good, we must suppose that this person may very well still have evil maxims and (perhaps a great many) after his resolution to be good. We might expect many of these to become foregrounded in his mind as the situations in which they are relevant arise since we are supposing he has made a *sincere* resolution to be good. Evil maxims are always marked by conscience and arguably one who is sincerely striving to be good will attend to this at least in

relation to some[100] of these maxims.[101] In addition, the agent who has come to acknowledge autonomy as freedom, is arguably better equipped than the evil person also to acknowledge licentiousness inherent in certain activities of his and to identify those habits which seem to have the appeal associated with this sort of freedom: the thrill of the exercise of power over others as opposed to the sober idea of power over oneself associated with autonomy; the demand of license that he 'respect' himself as much as he 'deserves' (and his 'desert' knows no limit according to license) as opposed to the command to respect everyone and himself equally as free and rational beings. There are no guarantees of recognizing false freedom because it is in the nature of licentiousness to disguise itself. But the fact that this evil is at the level of the maxim and he is *ex hypothesi* good at the level of the *meta-maxim* should stand him in good stead in the task of extirpation. It may also help him to resist new temptations and new evil maxims.[102]

Although, one might at first question the point of *the duty of moral perfection* this can be explained fairly easily. Arguably, since each of our particular duties, as a duty, already tells us that it should be done, we effectively already have the collective message from our total set of duties that we should do all of them. This seems to suggest that a special duty telling us to do all of our duties is superfluous. In response to this, we might say that whilst it is true that each duty tells us that it ought to be done - if only we would hearken to it - it is nevertheless conducive to more of them actually being done that there be an explicit duty to set an end of doing all of one's duties. In pursuing a duty of moral perfection, I am trying to overcome the difficulty of always being aware of what the law requires of me in a given situation. This seems to be particularly challenging in the case of duties of virtue (for example, the duty to promote the happiness of others), which are *imperfect* duties and as such do not specify a particular action. As regards the method of this duty, it seems to me that we must attend to the necessitating thought of the law in all situations.

The establishment of the remaining elements of moral development is less straightforward and it will require lengthy discussion to do this. Firstly, as regards *the duty to be holy*, it is not obvious why we should strive to make our motives absolutely pure as it commands us to do since arguably one may still do one's duty even if one acts from mixed (moral and non-

---

[100] Later we will see that he may not have the strength to confront them all.

[101] Although as a good person he identifies evil maxims through conscience more readily than an evil person, since the latter is still capable of attending to conscience, the good person should not take his frequent identification of evil maxims as indisputable proof of his goodness since such frequency provides no clear indication of it.

[102] Again, the good agent's greater ability to recognize licentious appeal for what it is would not constitute grounds for him to claim that he *knows* that he is good since there is a sense in which even the bad (dimly) recognize license it for what it is (or else they are not responsible): i.e., the difference between the good person and the evil one in this respect is not clear enough to the good person for him to be able to judge with certainty that he is good.

moral) motives. Secondly, there may be a clash between the requirements of this duty and *duties of cultivation* (the most important of which is arguably *the duty of humanity*). Conversely, there may not be a point to such duties of cultivation if the purification (the holiness) of maxims - the outright extirpation of empirical interests in maxims in accordance with duty - is in fact what is required. These issues (concerning purification and cultivation) are addressed in the next sub-section (3.1). Following this I would like to establish as an element of development *the agent's ability to reassure herself of moral progress*. Next, since the duties to be holy and of humanity as well as the pursuit of reassurance all require that the agent be capable of assessing the moral status of his maxims, it is necessary to examine the prospects of doing this.[103] This will complete the establishment of the elements of moral development (save virtue itself). We can then set about developing a deeper understanding of virtue itself in order to see how it might be acquired. As mentioned earlier, it will emerge that virtue is linked to the pursuit of all the other elements of development as well as many if not all of their intended end-states. Let us now turn to the aforementioned discussion of the duties to be holy and of humanity.

3.1 The duty to be holy and the duty of humanity

The duty to be holy requires that one strive towards an end-state of purity of motivation. In the *Doctrine of Virtue*, Kant describes the end of the duty to be holy thus:

> this perfection consists subjectively in the *purity* (*puritas moralis*) of one's disposition to duty, namely, in the law being by itself alone the incentive, even without the admixture of aims derived from sensibility, and in actions being done not only in conformity with duty but also *from duty*. (MS 6:446)

Presumably the type of maxim which is to be purified is that which is merely in accordance with duty since this is the only type which can be moral but might not be pure. Kant merely outlines the duty to be holy without justifying it, perhaps because he regards it as obvious that a moral being ought to strive for purity. One argument in favour of the project of purification is that firstly, it is only acting from duty which can guarantee that one is doing what is right. Secondly, whilst an empirical interest may co-operate with duty under certain circumstances, allowing it to remain as a ground of one's maxim is risky since it may actively go against duty under different circumstances. For example, when the customer is

---

[103] Strictly speaking, I do not require access to my motives in order to carry out the duty of moral perfection because it simply requires that I do *all* of my duties. That I do them *from duty* is the particular requirement of the duty to be holy. If it were the case that in addition to requiring that I do all of my duties, the duty of moral perfection also required I did them all *from duty*, it would make the duty to be holy redundant.

experienced, the profit motive may prompt me to charge a fair price but when he is inexperienced, the same motive may prompt me to overcharge. Duty, on the other hand, always commands me not to cheat the customer. (Herman, 1993) There is a more fundamental[104] argument from freedom in favour of purity: only by acting *purely* from respect for the law can I thoroughly affirm my autonomy which I, *qua* will, have overriding reason to do.

However, it may be that the requirement to strive for absolute purity is contradicted by another duty. In the *Doctrine of Virtue*, shortly after the duty to be holy, Kant explains that we have a *duty of beneficence* (or 'practical benevolence'[105] as he calls it) (MS 6:450-452) which is a command to help one's fellow men according to one's means. He then goes on to describe another duty, related to beneficence, which he calls the *duty of humanity*. This is a command to cultivate naturally occurring feelings of compassion (or sympathy). He says,

> *Sympathetic joy* and *sadness* (*sympathia moralis*) are sensible feelings of pleasure and displeasure . . . at another's state of joy or pain (shared feeling, sympathetic feeling). Nature has already implanted in human beings receptivity to these feelings. But to use this as a means to promoting active and rational benevolence is still a particular, though only a conditional,[106] duty. (MS 6:456)

As this passage continues, it is very unclear exactly what the aims of this cultivation are supposed to be. However, analysis of it raises the suspicion that it might conflict with the requirements of the duty to be holy, as we shall see shortly. In the quotation above, the initial wording of the command is to use sympathetic feeling 'as a means to promoting *active* and *rational* benevolence' (MS 6:456; emphasis added) He then expands upon this a little by distinguishing between 'the *capacity* and the *will* to *share in other's feelings* (*humanitas practica*)' and the mere '*receptivity*, given by nature itself, to the feeling of joy and sadness in common with others (*humanitas aesthetica*)' (MS 6:456) The first is free and based on practical reason; the second is unfree and is communicable (it seems, in a mechanical, unthinking fashion as it is likened to the spread of warmth or disease).

That the aim is to end up with a form of sympathy over which we have some sort of rational control - i.e., *humanitas practica* - is further indicated by his reference to 'the Stoic' who takes it to be wise to wish for a friend, not to have someone to help him should he need it but in order that he might help that friend, *yet* despite this, 'when he could not rescue his friend, said to himself "what is it to me?" In other words he rejected compassion.' (MS 6:457)

---

[104] This is a more fundamental reason than Herman's because the reason non-accidentally doing one's duty has (incomparable) value is that doing so is non-accidentally to affirm one's autonomy.

[105] What we might call 'ordinary' or 'non-practical' benevolence is 'satisfaction in the happiness (well-being) of others' (MS 6:452), without actually doing anything to contribute to their happiness.

[106] Kant offers no clue as to why it is conditional or what the condition is.

To show he supports this position (or attitude), Kant points out that where one person is in difficulty and I cannot help him, there cannot be a duty for me to suffer with him from compassion since this would only increase the ills of the world in a way which does not help him and, moreover, it is an affront to his dignity to be patronized in this way. In developing some sort of control over 'raw' compassion, we afford ourselves the *freedom* (or perhaps, rather, we make use of our pre-existing freedom) to avoid self-indulgent sentimentality. Of course, this 'negative' aspect of control in no way conflicts with the duty to be holy (if anything it supports it).

However, whilst Kant's pronouncements in the final paragraph of his exposition proper[107] of *humanitas practica* are rather vague, it is clear that this duty has *some positive* role to play even if it is unclear what this role might be. He says 'it is a duty to sympathize actively' in the fate of others and that it is 'an indirect duty to cultivate compassionate natural (aesthetic) feelings in us, and to make use of them as so many means to sympathy based on moral principles and the feeling appropriate to them' (MS 6:457). He then tells us we should also visit places such as debtors' prisons so as not to avoid sharing painful feelings 'For this is still one of the impulses that nature has implanted in us to do what the representation of duty alone might not accomplish.' (MS 6:457) This last claim (since it includes the word 'impulses') suggests the aim of cultivation has something to do with motivation but it is not decisive evidence of this. The potential clash mentioned above is that the duty to be holy tells us to omit all empirical motives from our maxims but there is a suggestion here that we need not do this after all. We shall return to this thread shortly.

In the meantime, despite Kant's giving no clear indication of the purpose of the cultivation of compassion in this paragraph,[108] Nancy Sherman, in her book, *Making a Necessity of Virtue*, takes it as textual evidence that one of the functions of a cultivated compassion is as a 'mode of attention' as she calls it. Briefly, this is a capacity to recognize those situations which demand moral action. If this is the intended purpose of cultivating compassion in the duty of humanity, then there is no hint of discrepancy between that duty and the duty to be holy since the former would not then be concerned with motivation as the latter is.

However, in addition to failing to provide adequate textual evidence, there are technical reservations associated with the notion of feeling (or 'emotion' as she calls it) being deployed as a mode of attention. Sherman says emotions can 'help us to track what is morally salient as morally salient in our circumstances, and thus locate possible moments for morally permissible and required actions.' (1997, p.145) She goes on to say that 'Sympathy, for

---

[107] By 'exposition proper' I mean the exposition of the duty up to the *casuistical question* on it.

[108] Or in the rest of his exposition of the duty of humanity.

example, draws us to occasions of distress or need' (ibid.). The idea is that 'We attend in a charged and alert way, taking in what detached reason or perception might miss' (ibid.). Sherman's language is sometimes ambiguous in this section[109] and may lead some readers to suppose that she thinks Kantian 'emotions' actually have a cognitive capacity. For example, she says situational information 'is often provided *through* the emotions.' (ibid., 146; emphasis added) And 'Emotions present information not in indefeasible ways' (ibid.). However, in Kant, feelings have no cognitive capacity since they are to do with the relation between the subject and object. They tell us nothing about the qualities of the object (see MS 6:211-212). Sherman is well aware of this but (rather unhelpfully) does not mention it in the section under discussion only revealing that she is aware of it much later (1997, pp.177-179) and in a different section.[110] Having said all of this, I can see no reason to deny that feelings such as sympathy may enhance the attentiveness of our (separate) cognitive capacities to potential moral situations. They may also help to record a deeper more vivid impression of what is morally important in the world and this may be what Kant has in mind in the injunction to go out into the world and experience its woes first-hand. If it is, then the duty of humanity does not conflict at all with the duty to be holy.

However, Sherman also argues that emotions, including compassion, could act as a moral *motive*. She says acting from duty has moral worth 'presumably' because being motivated by it non-accidentally issues in right action.[111] However, she supposes that if an emotion such as compassion could be cultivated by reason 'so that to act from it involved a concern for the rightness of the action' (ibid., p.150), then we could also non-accidentally do whatever duty demands but act *from* compassion. Again, she is making use of another of Herman's findings,[112] which is the notion of duty functioning as a 'limiting condition'. This is a function other than the more familiar, straightforward moral motive. As a limiting condition, duty restricts those empirical incentives which can be incorporated into a maxim. In other words, it is reason applying the categorical imperative test for permissibility.

Sherman does not appeal to the corpus in support of the claim that compassion could act as a moral motive. But whilst, as I say, there is no decisive evidence in the exposition of the duty of humanity indicating the precise end Kant has in mind, the characterisation of compassion as one of the '*impulses* that nature has implanted in us to do what the representation of duty alone might not accomplish' (MS 6:457; emphasis added) towards the

---

[109] Section 4 of Chapter 4 of *Making a Necessity of Virtue*.

[110] The discussion I have analysed is from Section 4 of Chapter 4 of *Making a Necessity of Virtue*. This new revelation does not appear until Section 7.

[111] This is a point I presume Sherman borrows from Herman's 'On the Value of Acting from Duty' (1993).

[112] Also from 'On the Value of Acting from Duty' (1993). Sherman acknowledges Herman's idea of duty as a limiting condition (1997, p.150n.73).

end of that passage arguably *suggests* that it is to be cultivated to be used as a motive. However, it should be noted that it is impossible to act from a rationally policed sympathy *alone*. If acting from sympathy is supposed to have moral worth in this situation, then it must be the case that when duty as limiting condition is deployed to check the permissibility of the action, it will have been discovered that the act is a duty: that it is obligatory (otherwise there is just no morally worthy act 'up for grabs' so to speak). However, according to Allison's response to a proposal[113] similar to Sherman's, if the act is obligatory, then the mere fact that it is obligatory

> is, of itself sufficient reason to perform it. Accordingly, if moral considerations are to govern at all in the case of obligatory actions, they must do so by serving as sufficient reasons, which is just to say that the action must be from duty. (Allison, 1990, p.120).

I take it that the thought is that *so long as* duty is deployed as limiting condition on empirical motives, it cannot be (or perhaps rather, just would not be) excluded from also acting as a motive in the case of a required action by such a conscientious agent. In such a case, sympathy is simply not needed as a motive. Therefore, it is still unclear what *positive* role sympathy has in moral agency. Another suggestion might be that although uncultivated compassion does not guide us to moral acts in the way duty does, by cultivating it we may channel it in the right direction and harness it to bolster an insufficient duty motive. The problem with this - again according to Allison (ibid., p.118) - is that it makes acting from duty conditional upon the presence of a non-moral interest which would seem to undermine the genuineness of the purported 'moral' part of the agent's motivation.

It seems, then, that there is no *real* motivational role available for a cultivated compassion over and above that which duty can do (and *would do* in *all* cases in which sympathy is cultivated). Regardless of what Kant was thinking when he outlined the duty of humanity, if it has any bearing on motivation at all, arguably the idea *should* be to cultivate compassion so that, in the event that a proposed act is unlawful, the agent constrains himself from duty to forbear, and where the act is required, he performs the act from duty. In the latter case, there may also be motivation from sympathy but this is morally harmless as it will already have been established that the act was required (and therefore also permissible). In short, the cultivation task ought to be to ensure empirical interests at least do not impede duty.

However, the original problem was that the duty to be holy seems to demand that we have *no* empirical interests in our moral maxims whereas the duty of humanity now seems to say that we *may*, so long as they are controlled by duty. One might argue that purification is unnecessarily stringent given that, as we saw, the will which is conscientious enough to check

[113] Put forward by Paul Benson (1987).

the permissibility of its empirical motives is also a will which will act *from duty* if it turns out an act is required - thereby ensuring that the demands of the law are non-accidentally met by a will which is affirming its freedom as autonomy - and this is after all what the advocate of purity is concerned about.

The question then arises of whether there is anything else to choose between rational cultivation on the one hand and absolute purification (even of motives that could be cultivated) on the other as far as the agent's ability to do his duty from duty is concerned. Arguably, an important difference between them is that it may be far easier to cultivate than to purify in the case of certain motives and certain maxims. One notable example would be a maxim to preserve one's life. In the *Groundwork*, Kant tells us that this is a duty but that most people do it anyway as a matter of course from an immediate inclination (G 4:397). Whilst it seems nigh on impossible to completely exclude the desire to live from a maxim of self-preservation (as the duty to be holy would seem to demand), it seems plausible that for certain morally strong people, it is within their capability to regulate this desire so as to make the choice of preserving their life *or not* in various situations dependent on the demands of duty at the time.

It may even be argued that if the concern is uncultivated desire and the risk of violations of the law (of licentiousness), then rational cultivation *can count* as purification since where it achieves its aim, every component of the agent's motivation is either respect for the law itself[114] or a desire constrained by it. Failing this and assuming the duty to be holy demands the absence of all empirical motives - even cultivated ones - the two duties (purification and cultivation) could mesh if we say that the ultimate aim is absolute purity but where this is difficult one must in the meantime at least regulate one's empirical motives. This seems the best way to ensure a will expresses its genuine freedom to the best of its ability.

However, it is vital that those maxims in accordance with duty whose grounds are merely empirical be at least cultivated (if not, actually purified so that duty is their only ground) since there is more at stake than just the non-accidental performance of duty or the expression of true freedom in individual acts. Where the agent has empirically grounded maxims in accordance with duty, there is the danger that such apparently innocent actions may mask a slide from acting from what might be called an innocent (lawful) self-love to acting from an evil and self-conceited attitude. It seems possible within Kant's strictures and those I have set out in this study that one could, for example, begin by carrying out beneficent deeds from a

---

[114] We saw above that successful cultivation of an empirical interest also introduces the motive of duty into the maxim. Cultivation therefore makes the maxim *impure* (since prior to cultivation, it will have been merely empirically grounded).

compassion which whilst uncultivated has no agenda but the gratification of the (lawful) desire to help others. Kant's philanthropist in the *Groundwork* is an example of this: we are told his actions have no moral worth but can still be laudable (G 4:398). However, if my arguments in Chapter 5 are correct, it is also possible to carry out actions which are outwardly the same (or at least very similar) to these but which differ inwardly in that they are meant to affirm a licentious conception of freedom and an inflated sense of self-worth. As mentioned above, the good person who has overcome an evil *Denkungsart* is better able to recognize licentiousness and the self-conceited narratives which he used to espouse for what they are but there is no guarantee of such recognition. In addition, that the outward actions in the licentious case and the non-licentious one might be indistinguishable, arguably makes such changes in motivation harder to detect than they might otherwise be. Recall from Chapter 4 that such similarity aids self-deception according to the Sartrean model. Eventually, the agent may even do away with quasi-moral actions more or less completely and lapse back into outwardly depraved actions. He may act from friendship or compassion or fellow-feeling but these must be regulated by reason not only to exclude violations of the law but also to ensure we do not use the actions associated with them to ease our way back into depravity.

We have established that the duty to be holy is justified but that in some circumstances, the agent may only be able to cultivate empirical motives. So far it seems these should both be included as elements of moral development. However, arguably both purification and the cultivation of empirical motives require that we have some sort of *access to our motives*. It is hard to see how one could carry out either task unless one had some idea that the empirical motive which requires attention exists. I would therefore like to examine the agent's prospects for gaining direct access to his maxims. I will do this shortly but first, I wish to argue for the inclusion of one more element of moral development, one which also requires us to be able to form at least fairly reliable beliefs about our motives: that element is the ability to reassure oneself of moral progress.

3.2 Reassurance of one's goodness: a further need for access to motives

It is perhaps natural to suppose that one who has been engaged in a process of moral development may be curious to know whether he is really has affirmed his commitment to morality, that is, whether he is making progress and is thus a good person. Kant - at least in the *Religion* - thinks that we cannot know for certain whether we are good or evil since we cannot have an 'immediate consciousness of the immutability of our disposition' (R 6:71) i.e.,

of our commitment.[115] He thinks the agent who wills to be good is well-advised to be cautious about ascribing goodness to himself. He says,

> one is never more easily deceived than in what promotes a good opinion of oneself. Moreover, it seems never advisable to be encouraged to such a state of confidence but much more beneficial (for morality) to "work out one's salvation with *fear* and *trembling*" (R 6:68)

Recall that in Section 2 we found that there is a difference between resolution and revolution and that the former does not guarantee or constitute the latter (any more than a promise can be kept merely by making it). We may suppose that there can be false resolutions in which the individual fails to commit to morality. Kant thinks there are cases such as this (R 6:68-69). But if I am right that in failing to commit, one also fails to adopt the moral meta-maxim, it must be the case that the person who makes an empty resolution at no point ceases taking license to be freedom or ceases having the evil meta-maxim. It is possible, then, that (in at least some of these cases) the agent's reasons for making a 'moral resolution' and then failing to commit to morality are sinister ones. Let us imagine one who has an overarching policy of license but has been fleetingly conscious of true freedom through a few isolated worthy acts. *Qua* evil person, he desires to maintain the evil policy which allows him to express freedom as he sees it. And now that he dimly suspects he is at risk of owning up to a conception of freedom whose exercise will impose an unwarranted constraint on his full expression of outer freedom, he may block this threat by fobbing himself off with a sham revolution. This he can do by 'vowing' henceforth to be a good person on the one hand but then failing to commit to goodness through moral progress on the other. Following this 'resolution', instead of trying to act from duty, he may engage in acts motivated by inclination but which, outwardly, may be indistinguishable from morally worthy ones such as those associated with the conceited pathological philanthropy discussed in the previous chapter. Since all of this begins with (an albeit fleeting) *consciousness of true freedom*, and is followed by an (empty) *resolution* and (merely outwardly) *good actions* which mimic a *commitment* to morality, the sham revolution may seem (to the one who wishes it were real) to bear all the hallmarks of the real one and thus may resemble it closely enough for him to be satisfied that it is genuine.

Since the adoption of the moral meta-maxim (i.e., being a good person) depends upon a commitment to morality and assuming one can fake this, one who aspires to be truly good (or anyone else for that matter) is unable to take the genuineness of her own commitment for granted. The good agent may have cause to be concerned that she is one of those people who indulges in pathological philanthropy and who is excessively confident about her moral status precisely *in order to conceal* it from herself. Once again, we can grant that the good

---

[115] This point is repeated on R 6:77.

person is better equipped to recognize the appeal of licentiousness for what it is since she has lifted her self-deception regarding freedom and is conscious of it from the 'outside' as it were. And again she should also be aware of the way she is liable to exploit similarities in order to deceive herself - similarities such as that between the true and false revolutions described in the previous paragraph. But *qua* good person she is also aware that this is no guarantee of goodness since there is no distinction between good individuals and evil ones in these respects sufficiently clear for the agent to be able to say that she is certain of her goodness. For example, *everyone* must be at least dimly aware of license *as* license or else those who submit to it would not be responsible for doing so. It may be that the mark of a good person as such is that she is less than completely confident in her commitment and in this way, ensures that she avoids the complacency of the licentious and morally self-conceited individual, and manages to stay committed and, thereby, to adopt the moral meta-maxim.

Whilst one must not be complacent, in the *Religion*, Kant also thinks that it is important for the good person to have some reassurance that she is good since, 'without *any* confidence in the disposition once acquired, perseverance in it would hardly be possible.' (R 6:68) This may be overstating it somewhat but we can see that there is some need for reassurance given the possibility of false revolutions. In both the second *Critique* and the *Religion*, Kant thinks the way to reassure oneself is to (somehow) focus on one's moral progress after making the resolution. He says, 'from the progress he has already made from the worse to morally better and from the immutable[116] resolution he has thereby come to *know*, he may hope for a further uninterrupted continuance of this progress' (KpV 5:123; emphasis added). It is unclear here exactly what the agent should be examining (his actions or his maxims, for example) to assess his progress. In the *Religion*, Kant is both clearer on the method - it is our deeds we should examine - and more cautious about its reliability, saying that one who

> has perceived the efficacy of these principles [of the good] on what he does, i.e. on the conduct of his life as it steadily improves . . . from that has cause to infer, but only by way of conjecture, a fundamental improvement in his disposition (R 6:68)

In the same passage, he also says, 'on the basis of what he has perceived in himself so far, he can legitimately assume that his disposition is fundamentally improved.' (R 6:68) Kant's caution about this method is well-founded for the obvious reason that outwardly good actions

---

[116] Here, Kant calls the sincere resolution 'immutable' (and earlier we saw it was 'unalterable' (R 6:48)). However, once again he cannot mean this literally - but instead must mean the resolution is *firm* - since, as noted earlier, it would otherwise make a fall impossible. Also, this time there is a further reason for thinking he cannot mean it literally: in the quotation given above he says that *knowing* the resolution is 'immutable' gives us *hope* of continued progress but it would hardly be mere hope if we *knew* the resolution to be *literally* immutable.

might be merely in accordance with duty but are entirely compatible with an evil disposition (R 6:30-31) and can even form part of self-deception ploy to ward off possible attempts at moral rebirth, as I have suggested above. Some might think, then, that this method of appraisal is literally worse than useless for these reasons (i.e., by Kant's *own standards*).

Perhaps there is one method of appraisal which does not merely consider outward action. Since an increase in moral strength is a form of moral development, it can be said to constitute commitment and makes one good. In the *Doctrine of Virtue*, Kant says that moral strength can be measured by the evil inclinations it overcomes (MS 6:394). Perhaps one could take overcoming of greater and greater evil inclinations as a sign of an increase in moral strength. However, this has its drawbacks: it is presumably easy to mis-remember the relative strengths of past inclinations or rather the strength of the lure of licentious freedom which recommended them. In addition, if the agent's temptations happen to be getting weaker, overcoming them will give no indication of whether or not he has gained in moral strength. Even if they are getting stronger and he overcomes them, he may have already been strong enough at the beginning to overcome the later, stronger ones. In fact, he may be regressing for all he knows. Even worse, to stand any chance of being useful, this method of appraisal relies on his being sure that he is resisting these evil temptations *from duty*. It is no indication of moral progress if in reality he is declining to act on them because of countervailing *empirical* incentives, as Kant himself points out, 'Very often he mistakes his own weakness, which counsels him against the venture of a misdeed, for virtue (which is the concept of strength)' (MS 6:392)

We saw in the previous sub-section (3.1) that the pursuit of the ends of the duties to be holy and of humanity require that we have some sort of access to maxims of action in accordance with duty and especially to any empirical grounds of adoption so that these may be eliminated or at least cultivated. We now see that in addition to this (assuming Kant is right that some reassurance of our goodness is needed if we are to remain in sufficiently good spirits to continue our moral project), it turns out that such access is also required to check progress in the purity and strength of the moral motive of those maxims in accordance with duty since arguably this will give us (a perhaps defeasible) indication of moral development and thereby a (similarly defeasible) indication of a commitment to morality and a hence of a good *Gesinnung*. Whilst outright evil maxims are also relevant to an evaluation of progress, these will always be marked by conscience (to which a good agent as such will presumably attend) and so the problem of access does not seem so pressing in their case.

The fact that the aspiration of moral self-knowledge (or at least, belief) is connected *both* with the need to check whether we are *good* (as we have seen in this sub-section) as well as

the duty to holiness and its concern that we are *pure* (as we saw in the last sub-section) is reflected in the duty of self-knowledge from the *Doctrine of Virtue*, (since this tells us not only to ensure we are good but also that we are pure). Kant says,

> This command is "*know* (scrutinize, fathom) *yourself*," . . . in terms of your moral perfection in relation to your duty. That is know your heart - whether it is good or evil, whether the source of your actions is pure or impure, and what can be imputed to you as belonging originally to the *substance* of a human being or as derived (acquired or developed) and belonging to your moral *condition.*
>
> Moral cognition of oneself . . . seeks to penetrate into the depths (the abyss) of one's heart which are quite difficult to fathom (MS 6:441)

It is interesting that this is a command to *know* one's moral disposition and the purity of the source of one's actions given that this requires knowing the grounds of one's maxims (and as regards the former perhaps even the meta-maxim). Anyone familiar with the secondary literature on Kant will know that there is a great deal of trepidation on the prospects of knowing our maxims. Let us now turn to this issue.

3.3 The prospects for cognitive access to maxims; using feeling as a guide

The two reasons I have suggested why an agent would be interested in gaining some idea of the grounds of her maxims are that either she wants to purify or cultivate those motives on the one hand or that she wants to check that she is making moral progress, on the other. If some sort of access to maxims is possible, then in practice, the agent may often (though not always) carry out both functions at once: checking to see what work needs to be done - what uncultivated, empirical motives remain - and at the same seeking reassurance that past work has been effective. These two functions may coincide inadvertently. Let us examine what Kant and some of his commentators say about the prospects for access to our maxims and their grounds.

In *Constructions of Reason*, Onora O'Neill tells us starkly that, 'Like other aspects of intelligible character . . . maxims are not objects of knowledge.' (1989, p.71) Allison improves on this rather gloomy outlook by arguing that if I cannot in some sense be aware of my maxim - and be aware of it as *mine* (or have the capacity to do these things), then it would not be a *principle* upon which *I* base my action - it would be nothing more than an 'unconscious drive or *habitus*' (1990, p.90). Nevertheless, 'I need not be explicitly aware of myself *as* acting on that principle' in order to act on it (ibid.). He thinks we can explicate our maxims through reflection but does not explain how.

However, since Kant thinks we *should* examine our maxims he presumably also thinks we *can* acquire the required information by doing so. In addition to the abovementioned command to know thyself from the *Doctrine of Virtue*, in the second *Critique* Kant also says,

> It is of the greatest importance in all moral appraisals to attend with the utmost exactness to the subjective principle of all maxims, so that all the morality of actions is placed in their necessity *from duty* and from respect for the law, not from love and liking for what the actions are to produce. (KpV 5:81)

Moreover, in addition to the fact (already mentioned) that the pursuit of the duties to be holy and the duty of humanity, seem to require access to our motives (at least), it would arguably make a nonsense of a categorical imperative which commands us to act according to a certain sort of *maxim* (rather than one, say, that commanded a certain sort of *action per se*) if we could not form some moderately reliable view of our principles and their grounds. However, it is unclear how reliable Kant thinks beliefs about our maxims can be. In the *Groundwork* he warns that, 'we can never, even by the most strenuous self-examination, get entirely behind our covert incentives.' (G 4:407) And he also says in this passage,

> it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty. (G 4:407)

Later, in the *Religion*, he expresses a similar thought saying,

> a human being's inner experience of himself does not allow him to fathom the depths of his heart as to be able to attain, through self-observation, an entirely reliable cognition of the basis of the maxims which he professes, and of their purity and stability. (R 6:63)

In the last two quotations, it is clear that Kant thinks our cognitive access is such that we can form beliefs about our maxims and their grounds of adoption but that these beliefs should be held with less than Cartesian certainty. It is also interesting how, in both of these, he discusses the issue in terms of *experience*, given that maxims are supposed to be merely intelligible. That was O'Neill's problem with the knowledge enterprise in the first place. The thought, then, might be that either beliefs about maxims are somehow possible without intuitions in the practical sphere or that we must *reconstruct* maxims from the available sensible materials, for example, through reflection on a past action. The former seems unlikely so let us examine the prospects for the latter: clearly, the physical movements I performed, or the words I uttered are straightforwardly available to knowledge. Similarly, the desires I experienced are empirical items, as is the memory of an end I may have visualised at the time. One may object that (whether the aim is purification-cultivation or reassurance

about one's disposition) this does not really allow us to say whether the maxim was pure since the presence of a desire does not mean it was acted upon.[117] However, in what is perhaps one of Kant's most optimistic pronouncements on this issue, he says in the second *Critique*,

> anything empirical that might slip into our maxims as a determining ground of the will *makes itself known* at once by the feeling of gratification or pain[118] that necessarily attaches to it insofar as it arouses desire (KpV 5:91-92)

One difficulty with exploiting this (either for the purpose of purification-cultivation or for that of seeking reassurance of one's commitment) lies in our ability to correctly associate the feelings Kant mentions with the correct action. For example, if I recognize that duty demands I help a certain stranger and I do so but then notice a feeling of gratification afterwards, I may believe that I helped her from inclination. But it could be that in reality I did it from duty and the agreeable feeling was left over from an earlier morally permissible but pleasurable action. If my reason for wanting to know my motivation is as part of a purification-cultivation project, then in such a case, I may wrongly believe I need to cultivate a desire to make it responsive to the requirements of duty (since although I suppose it led to an act in conformity with duty this time, I am aware that it might fail to do so in the future). Or if, instead, the reason I wish to know my motive is because I am interested in being reassured about my moral commitment at that moment, I may wrongly take this episode as evidence that I am not committed. Conversely, if we cannot reliably associate feelings (of gratification or pain) with the correct particular actions, an individual may tell herself that the present action has moral worth (even though it is merely in accordance with duty) and that the feeling of inclination-based pleasure she is experiencing emanates from a different action. She would then overlook a motive which required purification or cultivation or if interested in reassurance of her goodness, she would wrongly take this as grounds for supposing she was committed.[119] All of this vindicates Kant's view that we can form some beliefs about our maxims and their grounds but these are not entirely reliable. Nancy Sherman suggests two methods which, although still defeasible, may be more reliable and provide more detailed information than the method of using feeling as a guide just suggested. We turn to this account now.

---

[117] Put another way: it does not mean that it was incorporated into the maxim as its ground.

[118] I take it that Kant is thinking that gratification would follow success and pain would follow failure to attain the end in question.

[119] She *may* be committed to morality but such an action should obviously not be taken as grounds for believing she is.

3.4 Sherman's two approaches to self-knowledge

In her article, 'Wise Maxims/Wise Judging', Nancy Sherman recognizes that Kant thinks that the moral project involves as a central aim making duty the sufficient motive of our actions (i.e., that striving for holiness is a duty)[120] and that therefore there is 'moral point to the notion of a conscientious vigil of what our maxims are' (1993, p.45) *despite* the limitations on self-knowledge imposed by the mere intelligibility of maxims (the issue we have just explored). She proposes two ways in which we may glean some information about our motives which circumvent this obstacle. Briefly, the first is to attend to the features of a situation which seem salient to us. For convenience's sake, I wish to call this Sherman's *situational method*. The second is to attend to the feeling of joy a virtuous person is supposed to feel as part of the exercise of virtue. Let this be called Sherman's *moral aesthetic method*. Sherman claims that the possibility of self-deception limits the reliability of moral self-assessment using her two methods. However, her account lacks a self-deception *story*. I provide this account thereby vindicating this point. Sherman also fails to acknowledge how self-deception affects reliability to varying degrees depending on the agent's findings. I also address this problem before adopting her account.

Sherman's situational method rests on the idea that our values determine to some extent how we see a situation, the features of it that stand out and what action it seems to require. Our values (our interests) also determine what she calls 'strategic ends' (ibid., p.50). Sherman uses Kant's example of the shopkeeper from the *Groundwork* to illustrate this idea: he has the strategic end of not overcharging customers either because of the value of self-interest or because of the value of duty. Obviously, knowing his end does not by itself help him to know his values (i.e., what might be motivating him) since the same end could be prompted by both sorts of value - a point Sherman herself raises. Her response to this - and this is the key point - is that there is a connection between end and value such that when an agent deliberates about how to construe a situation and about what ends to pursue - i.e., what to do, 'she is often forced back to what she really cares about and how those values do or don't find expression in her choice of action. And bringing to bear those values often clarifies the possibilities for present action.' (ibid., pp.50-51) If I have helped someone in the past and I am curious about my motives, I should not simply ask myself why I did it. Instead, I may acquire some idea of the grounds of my action when the situation arises again and 'I ask, prescriptively, *what should I do* this time?' (ibid., p.51) since this may involve my asking myself 'how much does it matter to me, what sacrifices am I willing to make, would I do it even if there is nothing in it for me?' (ibid.) She admits that the approach is defeasible since,

---

[120] She also hints that reassurance of our goodness is a desideratum and hence another reason for trying to ascertain our motives but does not report the importance, which (as we have seen) Kant places on reassurance, in the *Religion*.

for example, one may still be vulnerable to self-deception with regard to the grounds which were actually active. But the point is that evaluation 'need not be conceived of as a moment of remote, theoretical introspection . . . It often is part and parcel of the question of what should I do, and thus arises, so to speak, on the battle line of action.' (ibid.)

Sherman gives a fuller example of the situational method which is worth examining since it illustrates how two different interests make different features of a complex situation salient during practical deliberation, (even where, as we shall see, the two interests suggest the same action and intention). Thus, it shows how the salience of some features rather than others might be used to glean some idea of one's motivation. She asks us to imagine a woman who is dying but does not know it and her daughter who knows and has to decide whether to tell her if she asks. Firstly, the considerations of benevolence may argue for lying since doing so will shield the mother from fear and anguish. But there is also the danger that having been deceived, she may discover the truth and if she does, then these adverse consequences will follow anyway *and* in addition, there may be a loss of trust in her daughter. Benevolence, then, may also recommend telling the truth.

However, she may also consider how she ought to respect her mother as a person. As Sherman rightly points out, Kant thinks the law demands we act in such a way that the other person can 'consent in or share in my own maxim of action' (ibid., p.54) since this is a right one possesses as a being capable of rational choice. But this is precluded by lying because a lie is not something to which the deceived person can consent. Therefore, respect for persons demands that we refrain from lying. Sherman points out that the moral way of thinking in the form of respect for persons and their agency leads to further considerations which are distinctive of that way of thinking. For example, we realise that by denying her the truth we deny her what she requires to make the right decisions in the time she has left, so lying violates her right to freely and effectively exercise her agency; it denies the due acknowledgement of the dignity of that agency. Another more subtle consideration is that lying will prevent her from being able to attach the sort of significance to the experiences she has in her final days which she would want to: she has, for example, a right to know that visits from friends are probably last visits.

We can see that although the empirical motive of compassion can recommend the same action as duty (i.e., to tell the mother the truth) and, in such a case, has the same immediate end (that she know the truth), the sorts of considerations emanating from the situation which enter into the agent's deliberations differ greatly depending on one's values. Seen from the points of view of different values, the situation itself seems to make different sets of demands: valuing compassion makes it seem as though the situation only demands the

limitation of pain and distress for a helpless loved one; valuing duty allows the agent to realise that the situation demands care but care consistent with respect for the person we see before us whose dignity we must preserve. In short, the way we see the situation is an indication of what interests we have incorporated into the maxim governing our response to the situation. An additional important point (and something which Sherman does not mention) is that this method can sometimes tell us whether an empirical interest is uncultivated: for example, the daughter's sympathy seems to be uncultivated because it recommends *lying* (as well as telling the truth). Unfortunately, however, if her sympathy only recommended telling the truth, this by itself would not be a very reliable indication of cultivation since both a cultivated and uncultivated sympathy could suggest this.

Sherman's idea is that we try to glean information about our interests from the kinds of thoughts we have about a situation when trying to decide what our end should be - about what to do. But one important possibility which I think she misses is that once an agent has asked himself all the questions he needs to ask to make a decision (and has made one) and in so doing has also gleaned some information about his interests, there is a possibility of going further than this and continuing to ask questions (even though the choice of end has been made) thereby accessing more specific information. For example, imagine that when dealing with a customer, the shopkeeper notices that he is aware that failing to charge a fair price may adversely affect his future business. This is enough reflection for him to decide on an end (of fair treatment) and he can also use this thought (about the prudence of fair treatment) as evidence that he has an empirical interest in not cheating his customers. However, so far it is not clear whether that motive is cultivated or not. He may, therefore, continue to ask himself questions to glean this (more specific) information even though he has already decided on an end (which is a natural cut-off point for deliberation). For example, he may ask himself whether the fact that this customer lives locally is a factor in his choice not to cheat him and whether he would do so if the customer were a tourist just passing through. However, as we shall see shortly, when the news is good, - when, for example, it seems one's motive is cultivated - it must be treated with caution because of the possibility of self-deception.

In her paper 'Moral Self-Knowledge in Kantian Ethics', Emer O'Hagan criticises situational method claiming that Sherman's agent is engaging in a form of Kantian *casuistry* but in trying to use it to ascertain her motives, she is using it for a purpose for which it was not intended and for which it is ill-suited. O'Hagan seems to be saying that the purpose of casuistry is to aid judgement in determining what the law demands in particular cases. Its approach is 'to open up a variety of maxims and contexts for consideration and to subject them to the categorical imperative' (2009, p.531). Quoting from the *Doctrine of Virtue* she reminds us

that Kantian casuistry 'is not so much a doctrine about how *to find* something but rather a practice in how *to seek* truth' (MS 6:411;[121] quoted in O'Hagan, 2009, p.531). She argues that a preoccupation with what I am really doing 'will potentially lead me to ignore salient alternatives'. (2009, p.531) The thought seems to be that if I am concerned with my own goodness I may too quickly decide that my action and I are good. I will then firstly, cease considering further alternatives which may contain the real moral option and secondly, may engage in justifying narratives for my choice of action. Kant, she rightly says, warns us time and again about the dangers of moral self-conceit.

It is not clear to me that the methods Sherman examines constitute casuistry since I take it that this is a process of thought experimentation that deals in hypothetical cases rather than the sort of real-life practical deliberation that Sherman has in mind. However, this may be a merely terminological issue. Nevertheless, although she is apparently unaware of it, O'Hagan's objection to seeking reassurance of one's goodness is inadvertently just as much a criticism of Kant as it is of Sherman since (as we have seen) he thinks it is necessary for agents to seek reassurance for the purpose of maintaining one's morale.

In any case, as I said in Sub-section 3.3, the agent who examines her motives for the purpose of identifying areas for improvement will in practice also be checking whether she has improved (in some cases whether she intended to or not). And Sherman's main concern so far as the situational method is concerned seems to be with agents owning up to empirical interests so that they can purify their maxims. But even if its results give rise to reassurance, then this in itself does not mean a good agent who uses the method is thereby obsessed about her moral purity to the detriment of actually managing to do the right thing as O'Hagan contends. Rather, I think that a good agent as such would exhaustively consider the alternatives, strive to make a morally good choice and *simply in doing the latter* would have available to her certain clues that indicate (defeasibly) what interests she has, what work still needs to be done and what progress she has made. These clues might include, for example, the fact that the situation was seen in a distinctively moral way (e.g., the daughter seeing her mother as *person* to whom she owes the truth). There is no reason to suppose that just knowing that this information can be regarded as an indication of an interest (moral or otherwise) will distract the good person from making the right choice since, presumably, the good person as such will regard making the right choice as the *priority* (and regard the need for reassurance that she is good as secondary). Of course, there are those for whom moral self-satisfaction is not merely primary but also the *only* interest and these people in deciding that their action was worthy or at least permissible will produce narratives to justify it as O'Hagan contends. But we should not think that a good person is at risk of becoming one of

---

[121] O'Hagan incorrectly gives the reference as MS 6:441.

these bad and self-satisfied people merely because she has a secondary interest in moral self-evaluation.

Moreover, if the situational method happens to indicate a moral interest in an action, the agent is still *capable* of not taking this as justifying certainty in her moral progress. Sherman explicitly points out that one must avoid assuming that situational information which may indicate a mere moral interest (as in the daughter's case) can also reliably tell us whether that interest was also the 'all-sufficient' *motive* since, apart from not having direct access to maxims, we may deceive ourselves that acts merely in accordance with duty were done from duty (1993, p.55).[122] The situational method, then, comes with a warning to beware of this. However, it should be noted that despite the fact that self-deception is thought to be the *main threat* to the reliability of the method, Sherman simply assumes its possibility (as does O'Hagan). Whilst I think it *is* possible to deceive oneself that a mere interest is also a fully-fledged motive and that agents should be wary of this, the possession of a suitable self-deception story allows us both to vindicate the claim that the possibility of self-deception affects the reliability of the method and to see more clearly the precise nature of the self-deceptive threats involved. Fortunately, we have a suitable account.

The first possibility of which the agent using the situational method should be aware is that even fundamentally good individuals can sometimes fail to live up to their commitment and give in to the incentive to license at the level of the maxim and may will a particular evil action. This means that even if one is fundamentally good, one may will the particular evil maxim of misconstruing the results of the situational method. When wondering why she has told her mother the truth, the daughter may exaggerate in her own mind a mere interest in morality contained in the maxim under investigation (the lawful maxim of telling the truth to her mother) so that this interest seems to be a (more impressive) all-sufficient moral motive. The former maxim (of exaggeration) is evil because it is one of disingenuously taking oneself to be better than one really is (i.e. of taking a mere moral interest to be a moral *motive*). That it is evil is important because in the account in Chapter 4, evil is required to mask the process of self-deception. Again, such deceptions are facilitated by a similarity: this time it is that between a mere interest and an interest that is also a motive. The limited[123] explanation

---

[122] It should be noted that if according to Sherman it is possible for an interest to make us see a situation in a certain way but that this does not guarantee that the interest is also a motive, then this point obviously presupposes a view of maxims which distinguishes between the notion of an interest which is also a motive and an interest which is not also a motive. This seems a plausible notion. It is also implicitly endorsed by Kant in his notion of *frailty* in the *Religion*. This is the idea that the agent may incorporate duty into his maxim yet find it insufficient to *move* him to action when the time comes (R 6:29).

[123] As I argued in Chapter 4, it is not possible to explain fully why one would do an evil thing such as this since such an explanation would be tantamount to a rational justification of irrationality.

for this possible evil act is that the person might need to believe she is better than she really is, perhaps as part of a wider ploy to *slide back* into depravity: it may 'cover her tracks', so to speak. The agent who finds that they are frequently tempted to take as a motive that which, on reflection, they are only entitled to take as an interest should be vigilant about a possible return to depravity.

The second possibility of which the good person using the situational method should be aware is much the same as the cautionary tale told in Sub-section 3.2 and (unlike the problem described immediately above) concerns evil at the level of the meta-maxim. The problem is that there is always the possibility that one is a fundamentally *evil* agent who is fleetingly conscious of isolated 'flashes' of true freedom (such as those associated with a *mere interest* in treating one's mother as a person). The evil person's ploy is to take these interests to represent more than they do (for example one might take a mere moral interest also to be a moral motive). This is done in order to convince oneself that one is affirming a revolution through moral progress (e.g., progress from a mere interest to a full motive) in order to distract oneself from the requirement to make a genuine resolution affirmed by *actual* progress. In short, when assessing evidence from the situational method, one should be aware that one might be a good person at risk of becoming an evil one (as described in the previous paragraph) or an evil person wishing to remain evil (as described in this) and in both cases one might be interested in making oneself out to be making moral progress by exploiting the similarity between a mere moral interest and a sufficient moral motive. Once again whilst the good person is better able to identify the expression of licentious freedom involved in such deceptions, she should not think her ability enables her to do this indefeasibly. She must take herself to be still vulnerable to self-deception when using the situational method and therefore should take the indication of a moral interest to be no more than that.

Sherman's worry about the reliability of this method for ascertaining motives (quite reasonably) concerns self-deception regarding our *goodness*. The motive is unclear in the example of the mother and daughter because in it, the daughter has more than one interest (including duty). Sherman does not mention this but perhaps the findings of the situational method are more reliable in cases in which there is no indication of a moral interest in the action but only of an inclinational one: one is hardly likely to deceive oneself into thinking that one acted from inclination alone (although, conceivably, if there is evidence that there is more than one empirical interest, one may deceive oneself into thinking that one particular empirical motive was active rather than another). Nevertheless, in cases in which one purported interest is duty (as in Sherman's mother and daughter example), she is right that

the possibility of self-deception makes the situational method inadequate in ascertaining the actual motive or motives in these.

However, she suggests that there is another method (what I have called the moral aesthetic method) which may be used to support the situational one in order to provide the agent with some indication of whether duty was the active and sufficient motive and perhaps the extent to which it approaches sufficiency. The method is to attend to a certain moral feeling: the joy the virtuous person experiences in the performance of their duty (i.e., in the exercise of virtue) since this indicates (albeit, defeasibly) the extent to which we are getting closer to purity in our maxims and thus what work there is still to do in the project of purification and may be grounds for feeling reassured. In claiming that there is such a feeling as this, Sherman is drawing on 'the famous note to Schiller' (ibid., p.58) in the early pages of the *Religion* in which the former concedes that 'A heart that is happy in the *performance* of its duty . . . is a mark of genuineness in the virtuous disposition.' (R 6:19n[124]; quoted in Sherman, 1993, p.58) To add to Sherman's reference from the *Religion*, we can also point to references to this feeling in the *Doctrine of Virtue*. For example, when discussing how we may turn a duty of right into a duty of virtue by doing what was owed from duty, Kant tells us that we thereby experience a feeling of 'moral pleasure that goes beyond mere contentment with oneself' (MS 6:391), which he calls *ethical reward* and is the feeling to which we refer in the saying 'virtue is its own reward.' (MS 6:391) Elsewhere in the same work (MS 6:484) he seems to be saying that joy in the performance of our duty is something we must aim for since without it, our virtue is not genuine. The hint (it seems to me) is that a grudging 'virtue' would be mere continence, like that of the Aristotelian *enkratês*, and therefore not virtue at all.

O'Hagan is concerned that the use of feeling as an indicator of 'a good heart' (2009, p.533) can go wrong. Unfortunately, she uses a wholly irrelevant example (borrowed from Richard Moran)[125] to illustrate this. Briefly, the example is of a man who treats the fact that he feels guilt over a bad action as a sign of his goodness and so feels good[126] (whereas, he should just feel guilty). However, the example is irrelevant because whilst it is possible to misuse our guilt in this way (to feel good about feeling bad), the moral feeling which Sherman is thinking of - moral *joy* - is purported only to occur when one has acted virtuously. There does not

---

[124] Sherman is relying on the edition of the *Religion* translated by T.M. Greene and H.H. Hudson and published by Harper (1934). The reference she gives for this quotation (and another similar one) is R 6:19n. This page (p.19) has been edited out of the Cambridge edition of the *Religion* but the two comments on joy in the exercise of virtue which Sherman quotes appear in the latter edition at R 6:25n.

[125] Moran, R. (2001) *Authority and Estrangement: an Essay on Self-Knowledge*. Princeton: Princeton University Press.

[126] And then feels bad about feeling good in a quasi-morally self-indulgent way.

seem to be any room for the same sort of abuse in the case of moral joy as was highlighted in the case of guilt.

A more serious worry (one which Sherman raises herself) is that one would not be able to use this feeling as an indicator of the purity of a maxim at all if it were impossible to trace the provenance of it to motivation by duty. However, she thinks that this is eminently possible since more than being just 'an external label signalling one's internal state . . . those feelings are themselves part of the subjective experience of being moved by the moral law, a part of what it is to be in the grip of its authority at that moment.' (1993, p.59) Sherman thinks that when used to confirm what we have learnt (or suspect) about our motivation using the situational method, the feeling of joy in acting virtuously gives us 'an alertness that is missing in the blunter grasp of a merely cognitive report.' (ibid.)

However, Sherman has her own doubts about the moral aesthetic method *since, again, she* thinks we are just as liable to self-deception with regard to feeling as we are with regard to 'merely cognitive' situational *reports. She does* not elaborate on this but we can imagine how (the intimate connection between moral joy and moral motivation *notwithstanding), a morally* self-conceited individual of the sort we met in Chapter 5, determined to convince himself that he is virtuous, could, for example, take the pleasant feeling he experiences having helped a person - from say, sympathy and from a licentious desire to dominate - to be the feeling of joy in the exercise of virtue. As we saw in Chapter 4, *similarity* between a truth and a falsehood which we want to believe aids self-deception and a virtuous act and a morally self-conceited one are both exercises (of different sorts) of power: virtue being the power to make the moral choice sometimes despite countervailing temptation and morally self-conceited action involving power over another human being. As in the case of the situational method, the good agent should not suppose that his greater ability to recognize the licentiousness of such deceptions enables him to do so indefeasibly and so he should continue to think that the possibility of self-deception regarding moral joy is a real threat to the reliability of this method. Sherman's worry about its reliability is thus vindicated. However, despite the spectre of self-deception, she maintains the method can still give us some idea of our ultimate values.

I agree but would add that there is one further problem with the reliability of the moral aesthetic method. If the agent is experiencing other feelings and these are strong enough to 'drown out' any moral joy which *might* exist to be felt, then the agent cannot be sure whether there is any moral joy to be felt (and we may expect other feelings often to occur in situations which demand the exercise of virtue). When the daughter tells her mother she is dying, it may be that she *would have* felt some joy at having treated her mother with respect

in this way if it were not for the overwhelming sadness involved in doing this. When other (strong) feelings are present, the wise agent will not draw any conclusion from the apparent absence of joy. The good but unwise agent may draw the wrong conclusion that he is less virtuous or pure than he really is and that he has more work to do than he really does to make the moral law his sole or at least sufficient motive. This may both draw his energies away from where they are needed - i.e., from maxims which *do* need improving - and may also adversely affect his morale.

3.5 Viability of the duty to be holy, of the duty of humanity and of moral reassurance

We have seen that at least three components of moral development: the duty to be holy, the duty to humanity and the ability to reassure oneself of moral progress rely on our having some sort of access to our motivations. Let us sum up how reliable Sherman's methods are and how effectively the agent can carry out the three functions which depend on the motivational information. Firstly, as regards pursuing holiness or cultivation, the situational method can indicate to the agent the presence of a moral or an empirical interest (alone) with a good degree of reliability. We have also seen that where an empirical interest prompts a violation of the law, it is safe to say that it is uncultivated but where it prompts something in accordance with the law, one cannot tell. Also, I think one may extend reflection beyond the point of choosing one's end in an effort to glean more exact information about one's interests. Finally, where the situational method tells the agent she has both an empirical and a moral interest in an action (as in Sherman's example of the daughter), it cannot tell her which one (if either one) is the sufficient motive if she performs that action. Where she apparently only has one interest, if it is duty, she might be deceiving herself (and my account of self-deception vindicates this claim) but if it is not duty it is arguably much more likely to be her sufficient motive (it could conceivably be another empirical interest which she has failed to foreground and acknowledge, which is the sufficient motive).

The moral aesthetic method is meant to help ascertain whether a moral interest we detect through the situational method is also a sufficient motive and perhaps the degree of purity of my motivation. However, we have seen that when I seem to be experiencing moral joy, (when it is good news), there is always the risk of self-deception. (This claim is again vindicated by the self-deception story from Chapter 4 but also this time by the account of moral self-conceit from Chapter 5.) In addition, the method is unusable if there are strong feelings to mask any moral joy that might be felt. It seems the only situation in which it is very reliable is when there is no other feeling to mask it and it is actually absent, indicating that motivation is (probably entirely) inclinational.

In short, the agent can use these methods as a fairly reliable indicator of empirical interests and if he asks the right questions can arguably quite often tell if his empirical interest is cultivated or not. Crucially, this means, that the opacity of maxims notwithstanding, we can at last say that *the duties to be holy and of humanity are viable*. However, mainly because of the doubt introduced by the possibility of self-deception when the news is apparently good, these methods are much less reliable in those instances, and thus the agent hoping for reassurance will simply have to make do with little reliable reassurance in the latter stages of development. Still, since self-deception is only a possibility, the good agent as such can draw *hope* from good news but no more than this. These findings regarding the reliability of Sherman's methods vindicate my earlier point that one must remain wary of one's goodness and the good agent as such is wary of it understanding the dangers of credulity when self-deception is a possibility.

4. The duty to acquire virtue (moral strength)

At the beginning of Section 3, I said that the core elements of moral development are the duty to uproot evil maxims, the duty to be holy, the duty to cultivate those empirical motives which can co-operate with duty (especially compassion), the duty of moral perfection, the ability to reassure oneself of one's moral progress and the duty to acquire virtue. We have established that those which require at least a defeasible access to motives for the purpose of diagnosing further problems to be addressed (the duties of purification-cultivation) are possible. The acquisition of virtue has, until now, been understood simply as the strength to do one's duty more readily than one could without it. I would now like to try to give a more sustained account of it. In better understanding what it is, we may see more clearly what the agent can do to acquire it.

Since *autocracy* - the capacity for rational constraint - is central (and perhaps identical) to virtue, understanding it may allow us to make headway. I would like to begin by presenting Anne Margaret Baxley's account given in her paper, 'Autocracy and Autonomy', which sharply distinguishes between these two concepts. This picture is then contrasted with Stephen Engstrom's view in 'The Inner Freedom of Virtue' in which he argues that autonomy and virtue are different but *linked* by the concept of inner freedom. It will emerge that Engstrom's account not only gives a more accurate rendering of the relation but also, in explicating the nature of virtue as inner freedom, it yields some insights into actual methods for the pursuit of virtue which cohere very well with the account of the initiation of the revolution (through consciousness of autonomy as true freedom), given at the beginning of this chapter.

4.1 Baxley on autocracy and autonomy

Baxley opens by giving us a preliminary sense of the relation between autocracy and virtue. She says Kant describes virtue as (amongst other things) 'strength of mind', 'courage' or 'fortitude' (2003, p.2) Autocracy is said to be the self-mastery or self-constraint central to virtue and is both a necessary and sufficient condition of it. She thinks understanding autocracy may illuminate virtue. The strategy for the former is to compare autocracy and autonomy and to do this she compares the moral capabilities of an infinite holy will (i.e., God), a finite holy will (Christ) and the finite and less than holy will of man. First, she quotes from the *Doctrine of Virtue* where Kant tells us that

> For finite holy beings (who could never be tempted to violate duty) there would be no doctrine of virtue but only a doctrine of morals, since the latter is autonomy of pure practical reason whereas the former is also *autocracy* of practical reason, that is, it involves consciousness of the capacity to master one's inclinations when they rebel against the law, a capacity which, though not directly perceived, is yet rightly inferred from the moral categorical imperative. (MS 6:383; quoted in Baxley, 2003, p.4)

Baxley says that where Kant compares virtue and holiness, the sort of holy will he typically has in mind is the infinite or divine will. She thinks the reason this will is immune from transgression is simply that it does not have a sensible nature and cannot be tempted. This is also why the moral law does not take the form of an imperative for such a will. However, a merely finite rational being (by which Baxley means a human being and not the finite holy will) has both a rational and sensible nature and has counter-moral inclinations. It is because of these that we must be *constrained* to obey the law and this is why the law presents itself to us as duty. The finite holy being considered in the quotation shares a feature with us: its finitude - by which is meant it experiences empirical inclinations - but it is also holy - which means it always resists these and so its actions never fail to accord with the law.

However, Baxley thinks the following considerations give rise to a potential problem for Kant: it was the absence of a sensible nature which made transgression by the infinite will inconceivable. One might think, then, that the possession of a sensible nature 'is sufficient for raising the specter of contra-moral action.' (Baxley, 2003, p.5) Given that the finite holy will has such a nature - i.e., that it experiences temptation - it seems Kant is not entitled to claim that it is *inconceivable* that it would ever violate the law. The answer - as we might expect - is that we think of such a being as 'afflicted by just the same needs and hence also the same sufferings' (R 6:64; quoted in Baxley, 2003, p.5) but also as 'superhuman, inasmuch as his unchanging purity of will, not gained through effort but innate, would render any transgression on his part absolutely impossible' (R 6:64; quoted in Baxley, 2003, p.5)

Although he is tempted, he is 'constitutionally incapable of succumbing to temptation.' (Baxley, 2003, p.6)

Baxley then poses three questions which she thinks the passage (MS 6:383) quoted above raises. The answers to these will reveal more about the nature of autocracy and its relation to autonomy. Firstly, we might wonder why the finite holy will is not in need of either a doctrine of virtue (a set of ethical duties) or of the autocracy of pure practical reason. Secondly, she asks why a doctrine of morals, which is connected to the autonomy of pure practical reason, would pertain to this finite holy being. And thirdly, she wonders what the passage tells us about the relation between autonomy and autocracy (ibid., p.7).

The answer to the first question is that although inclinations occur to such a will and these can themselves sometimes be counter-moral ones, it never takes them as a reason for action since it uniformly prioritizes duty in every choice. This reveals the more interesting point that not only does virtue presuppose counter-moral inclinations but *also* the mere *possibility* of the will taking them as reasons for action. Baxley believes that this is the reason Kant thinks that although virtue 'signifies' moral strength of will, it does not exhaust the concept: *strength can also be attributed to a holy superhuman being* 'in whom no hindering impulses would impede the law of its will' (MS 6:405). For Kant, this leads to the point that, 'Virtue is, therefore, the moral strength of a *human being's* will in fulfilling his *duty*, a moral *constraint* through his own lawgiving reason, insofar as this constitutes itself an authority *executing* the law' (MS 6:405; quoted in Baxley, 2003, p.7). The thought seems to be that our susceptibility to violating the law means that virtue is a particular kind of strength only we have (and need to develop): it is strength through rational constraint.

To answer the second question, Baxley first draws on the *Groundwork* for a definition of autonomy where it is said to be a property of the will 'by which it is a law to itself (independently of any property of the objects of volition)' (G 4:440; quoted in Baxley, 2003, p.8). She takes this to mean that autonomy is a will's capacity to 'impose principles on itself that make no reference to a state of affairs an agent desires to bring about' (2003, p.9). The finite holy will is autonomous in this sense of being capable of giving itself laws which are 'not directed toward satisfying any interest in an object of volition.' (ibid.) Since a finite holy will legislates the law and follows it perfectly, the sense in which a doctrine of morals is applicable to it is that it is merely descriptive of what it does rather than a set of duties for it.

Baxley is now is a position to answer the third of her questions. She asked what the above-quoted passage[127] tells us about the relationship between autonomy and autocracy. Her first

[127] (MS 6:383).

point is that although not all autonomous wills are obliged to become autocratic, autonomy is a necessary condition of autocracy since 'only a will capable of self-legislation in accordance with pure practical reason can acquire the requisite strength to have pure practical reason consistently determine the will, even in the face of countervailing inclinations.' (ibid., p.10) She then draws on the *Lectures on Ethics* where Kant is reported to have defined autocracy as 'The power that the soul has over all faculties and one's entire condition' and to have stated that 'If a man does not strive for this autocracy, then he is a plaything of other forces' (VE 27:175; quoted in Baxley, 2003, p.10). Autocracy, then, involves disciplining one's sensible nature - or as she later says, more precisely, it involves disciplining ourselves in the way we value sensible inclination. Kant then sets out two different types of authority. Quoting from the same passage, Baxley finds that,

> Our authority over ourselves is both productive and disciplinary. As executive authority it can, in spite of every obstacle, compel us to produce certain effects, in which event it has might. As directing authority it can only guide the forces of character. (VE 27:175-176; quoted in Baxley, 2003, p.10)

According to Baxley the thought is that productive authority is an 'executive' or 'compulsive' one whereas disciplinary authority is 'legislative', 'directing' or 'guiding'. The latter is the capacity to provide rules and the former is the capacity to govern in accordance with them in spite of counter-moral incentives. That the executive authority is autocracy Baxley thinks is shown in the quotation from the same passage in the *Lectures on Ethics*, 'Autocracy, therefore, is the power to compel the heart in spite of every obstacle. Mastery over oneself, and not merely directing authority, belongs to autocracy.' (VE 27:176; quoted in Baxley, 2003, p.11) And a further quotation from *Moral Mrongovius* II she thinks shows that again that it is autocracy that is the executive authority and also that autonomy that is the disciplinary one (providing the rules), 'When reason determines the will through the moral law, then it has the power of an incentive, then it has not merely autonomy but also autocracy. It has legislative and also executive power.' (29:626; quoted in Baxley, 2003, p.11) Baxley clarifies her findings by stating that as autonomous beings we not only provide the rules for action but also 'possess the requisite freedom of will to act out of respect for the law alone.' (2003, p.11) However, she says, a human being requires autocracy to act consistently from duty because she must master her sensuous nature in order to do so (ibid., p.12). Virtue does not consist simply in a capacity for autonomous willing; it is a capacity for autonomous willing in the face of a susceptibility to value countervailing incentives.

Baxley's account provides us with some insight into virtue, for example, by explicating the notion that the capacity to constrain is only relevant for 'merely finite' rational beings such as ourselves. In addition, we must concede that there is textual evidence that suggests that

Kant draws a distinction between autonomy as a legislative authority and autocracy as executive authority. However, Baxley's account entirely ignores the notion of inner freedom. As we shall see, this oversight leads to a failure to consider counterarguments to her overly-sharp distinction between autonomy and autocracy. Let us now see how Engstrom's account deals with these issues.

4.2 Engstrom on inner freedom as capacity and inner freedom as virtue

Engstrom's stated main aim is to explore the conception of freedom prevalent in the Introduction to the *Doctrine of Virtue*. In so doing, he will dispel the view that Kantian virtue amounts to mere continence (Engstrom, 2002, p.291). However, as I mentioned at the beginning of this section, the main value of his account is that it provides a reading of virtue as inner freedom, which firstly, I take to be correct and secondly, coheres with the account I have given of the initial resolution to be good through consciousness of true freedom.

He begins by noting that Kant rejects the notion that virtue is custom or habit because if it were, it would not be armed for all eventualities as it needs to be (MS 6:383-384). In addition, as Felicitas Munzel points out, principles passively held (as a habit is held) 'could not meet the requirements of resolve, that is, of conviction and consistent adherence' (1999, p.226). Engstrom tells us that Kant's further reason for rejecting the notion that virtue can be understood in terms of custom is that in grounding maxims on custom the agent would 'lose *freedom* in adopting his maxims, which [freedom] however is the character of an action done from duty' (MS 6:409; quoted in Engstrom, 2002, p.293). Engstrom wonders whether the thought is that custom involves falling into a rut in which choosing otherwise becomes more difficult and that this might erode our freedom in choosing maxims since such freedom (seems to) imply the possibility of choosing otherwise. On reflection, this is rejected because (as we saw in Chapter 3) Kant denies that freedom entails the ability to choose otherwise. This can be seen in the claim that 'The less a human being can be compelled physically [that is, coerced], and the more he can be compelled morally (through the mere representation of duty), the freer he is' (MS 6: 382n; quoted in Engstrom, 2002, p.293). One who is compelled morally is free by virtue of that very compulsion and yet if so compelled is not capable of choosing otherwise any more than one who is following custom.

Alternatively, then, Engstrom suggests one might think that if freedom consists in moral compulsion and if custom limits our freedom then,

> the choice of maxims on the basis of custom involves a loss of freedom in adopting maxims precisely because such a basis limits the extent to which we can be compelled morally, in so far as

> such compulsion involves quite a different basis for choice - namely the mere representation of duty. (2002, p.293)

The notion that custom limits freedom as *moral compulsion* might seem to be equivalent to the notion that custom involves a loss of autonomy.[128] However, this is rejected since grounding a maxim on custom in no way affects the subject's *capacity* for acting on the moral law: that is, it does not affect her practical freedom either negatively or positively understood - we have no reason to suppose that the agent acting from custom would thereby lose his accountability. If Kant thinks (some type of) freedom can be limited (by custom for example), but practical freedom and one's accountability cannot be limited, then, Engstrom concludes, Kant must have another conception of freedom in mind in his claim that custom restricts it.

Engstrom proposes that this is *inner freedom*, a conception of freedom which Kant associates with *virtue* but not with *accountability*. However it is not clear how it is different from practical freedom. The most sustained account of inner freedom is in *The Metaphysics of Morals*, but Engstrom begins by drawing on the brief account in the second *Critique* where Kant tells us it is a 'capacity [*Vermögen*] . . . to release oneself from the impetuous importunity of the inclinations to such an extent that none of them, not even the dearest, has influence on a resolution for which we are now to make use of our reason' (KpV 5:161; quoted in Engstrom, 2002, p.298)

Engstrom thinks some might take this to be similar to negative freedom. But he points out that the independence from the *influence* of inclinations on choice which inner freedom affords us is not the same as the independence from brute *determination* by sensuous impulses afforded us by negative freedom. Quoting the *Doctrine of Virtue* (MS 6:213), he reminds us that Kant takes the human will to be sensuously *affected* but *not determined* by sensuous impulses; that is, it is *sensible* yet *free*. But, he says, influence does not figure in Kant's descriptions of the 'general relation' of the power of choice to sensible impulses and one may or may not be influenced by an impulse. Quoting from the second *Critique* (KpV 5:118), he says 'while human freedom does not imply "complete independence from inclinations and needs", it is still possible "to hold the determination of one's will free of their influence"' (2002, p.299). Inner freedom's independence from the influence of sensuous impulses seems to be something beyond the independence from brute determination of negative freedom. Engstrom hints at the notion that the former is an independence from temptation.

---

[128] We cannot claim that at MS 6:409 Kant means we are less free in the sense of *acting* less autonomously when we adopt a maxim from custom since, in this passage, he is discussing the freedom *with which we adopt a maxim* rather than how freely a given adopted maxim allows us to act.

We may think then that we have here a conception of freedom associated with virtue, rather than the sort of freedom which makes us accountable and that therefore the difficulties associated with Kant's claim that custom robs us of our freedom are removed. However, Engstrom (ibid., p.300) thinks the following considerations show that inner freedom may not be any more closely associated with virtue than it is with practical freedom (or rather, autonomy as he later says, (ibid., p.301)): since the latter is a capacity to choose according to the moral law, it must also involve a capacity to be independent of the *influence* of sensible impulses (even if one fails to achieve it).

In addition, that we have a capacity for independence from influence by sensible impulses by virtue of our autonomy alone he thinks is shown in *the consciousness of obligation*: this consciousness, he says, has two aspects: the first is an 'intellectual representation of reason' (Engstrom, 2002, p.300). Reason represents to one 'the thought of oneself as *necessarily* acting in accordance with it [the law] so that the possibility (though not of course the bare conceivability) of one's acting otherwise is excluded.' (ibid.) The second aspect is the feeling of being necessitated by the law which necessitation implies constraint in the face of a sensible nature which may occasion reluctance. This is an awareness of 'affection by sensible impulses' (ibid., p.301) and hence an awareness that the fact that one might allow impulses to influence one cannot be excluded.

However, it is the modality of the intellectual representation (the first aspect of the consciousness of obligation) that is important in showing that autonomy involves a capacity to be independent of the influence of impulses. The argument runs as follows: the thought that necessitates in the consciousness of obligation is a practical representation of oneself as necessarily acting in conformity with the moral law. This is because firstly, this modality is 'internal' to this necessitating thought and secondly, because the idea of autonomy (revealed in the consciousness of obligation) is the idea of the capacity to *realise* what is represented in the necessitating thought. This means that the idea of autonomy 'contains from the start the idea of the capacity to exercise the power of choice in such a way that through this exercise one *necessarily* acts in conformity with the moral law.' (ibid.) One can only choose in such a way that necessarily accords with the law if choice is exercised in such a way as to secure independence from influence by sensible impulses. Therefore, 'the original idea of freedom is from the start the idea of a capacity to release oneself from the importunity of the inclinations to such an extent that none of them has influence on the power of choice.' (ibid.) He thinks this is why 'Kant says, the "consciousness of the *capacity* to become master of one's inclinations that rebel against the law" is, "though not immediately perceived, nevertheless correctly inferred from the moral Categorical Imperative"' (MS 6:383; quoted in Engstrom, 2002, p.301).

He concludes that the capacity Kant characterizes as inner freedom is a capacity we already possess in our autonomy. This is something Baxley denied[129] in claiming that autonomy only provides 'legislative authority'. However, in claiming this she did not consider the involvement of inner freedom in autonomy. Returning to Engstrom's argument: he says that if we possess a capacity for inner freedom by virtue of our autonomy, possession of it does not imply virtue any more than our autonomy does. However, whilst the notion of independence from influence does not pick out virtue, it is a feature shared by virtue and autonomy. Still, the relation between inner freedom and virtue remains unclear. To better illuminate this, Engstrom turns to the lengthier treatment of inner freedom given in the *Doctrine of Virtue*.

Drawing from that work, he says inner freedom can be defined negatively as 'independence of the power of choice from inner sources' (2002, pp.302) - i.e., sensible impulses - in the determination of one's will (and in action) (ibid., pp.302-303). Positively, it is 'the capacity [*Vermögen*] for self-constraint not by means of other inclinations, but by pure practical reason' (MS 6:396; quoted in Engstrom, 2002, p.303) Engstrom combines this notion of self-constraint with that of releasing oneself from the *influence* of sensible impulses (from the earlier definition from the second *Critique*) and supposes (quite reasonably) that the idea is to release oneself from the said influence by means of the said self-constraint. However, this new characterisation again fails to pick out a type of freedom peculiar to virtue: 'Just as having the capacity to release oneself from the influence of the inclinations does not imply that one has released oneself, so having the capacity for self-constraint does not imply that one has constrained oneself.' (2002, pp.303-304) The breakthrough is provided by the following passage from the *Doctrine of Virtue*,

> One may also indeed say that the human being is obliged *to acquire* virtue (as a moral strength). For while the capacity [*Vermögen*] (*facultas*) to overcome all sensibly opposing impulses can and must be absolutely *presupposed* on account of his freedom, yet this capacity as *strength* (*robur*) is something that must be acquired (MS 6:397; quoted in Engstrom, 2002, p.304).

Engstrom argues that bearing in mind his earlier analysis of inner freedom (as a capacity to release oneself from the influence of inclination through rational constraint), it is clear that the capacity to overcome all sensible impulses, presupposed 'on account of freedom' (mentioned in the present passage), is just this inner freedom. This passage links this capacity to overcome all sensible impulses to virtue as moral strength and since the former is inner freedom, it links inner freedom to said virtue.

The claim, then, is that there are two different but related senses of inner freedom at work here. The first is a sheer capacity to overcome all sensible impulses through rational

---

[129] Although, it is Allison's rendition of it which she rejects. (She does not acknowledge Engstrom's essay at all.)

constraint. In this sense, a capacity for inner freedom does not admit of degrees: it is something a being either has or does not have. We all have it in virtue of our practical freedom (or, more specifically, our autonomy). The second sense is of a developed form of this capacity: hence it is called *strength* (where 'strength' means the state of being stronger than those who have not engaged in the development of their inner freedom). It seems to me that inner freedom in the second sense *does* admit of degrees.[130] Engstrom takes it that the distinction between *inner freedom as capacity* which the agent has in virtue of her autonomy, and *inner freedom as strength* allows us to see how she may lack freedom in the second sense and yet retain freedom in the first sense and hence never fail to be accountable.

4.3 Engstrom on inner freedom as a moral promptitude and as combative of affects and passions

Engstrom then begins to examine inner freedom as a strength that must be acquired. We have seen that Kant rejects the notion of virtue as habit or custom. However, Engstrom reminds us that in the *Doctrine of Virtue*, Kant considers a notion of habit (*habitus*) as 'readiness' (*Fertigkeit*)[131] and then having characterised habit as a 'promptness to act'[132] he acknowledges that virtue is a particular kind of habit which can be characterised as a 'free readiness'. (Engstrom, 2002, p.306) This contrasts with the habit of custom which is a 'uniformity of action that has become a *necessity* through its frequent repetition' (MS 6:407; quoted in Engstrom, 2002, p.306), whereas free habit proceeds from freedom. Engstrom thinks the sort of freedom from which this free habit proceeds is inner freedom as capacity but that, as a developed promptitude or readiness, it is inner freedom as *strength*. The readiness or promptitude is not to be placed in the actions themselves or even in the power of choice but rather in the capability of freedom itself. He thinks this is shown in the definition of the *developed* capacity of freedom, 'readiness in free lawful actions to determine oneself through the representation of law in action' (MS 6:407; quoted in Engstrom, 2002, p.307). He therefore finally defines virtue thus:

> virtue is freedom ready to exercise itself; it is the *capacity* to determine oneself through the representation of the law in action in so far as it has developed into a readiness of freedom, which makes for facility in its exercise. (2002, p.307)

---

[130] I thank Seiriol Morgan for this useful way of seeing the distinction between inner freedom as capacity and inner freedom as strength. (Personal correspondence.)

[131] Gregor translates this as 'aptitude' in the Cambridge Edition (1996).

[132] This is Engtrom's own translation of *Leichtigkeit zu handeln* (2002, p.306). He prefers this to Gregor's rendition 'facility in action' partly because his own rendition preserves the infinitival construction *zu handeln* which, he says, indicates Kant is speaking of a quality belonging to a *capacity* rather than to its *exercise* (ibid., p.306n.).

For much of Engstrom's exposition, it seems that he takes virtue to be a developed capacity for self-constraint and hence independence from even the influence of inclinations in *the choice of maxims*. However, in the final section of the paper, his focus on the idea of virtue as a free habit emphasises the promptitude of freedom in determining itself to *acting* on already adopted maxims. In short, (although Engstrom does not seem to be aware of the difference) there seem to be two applications of virtue as inner freedom: in moral maxim *adoption* and in managing to *act* on a moral maxim. Kant himself refers to virtue in both of these ways. Although there are far fewer references to the former application of virtue, it is evident in a key passage in the *Doctrine of Virtue* in which Kant explains how one may make the right of another *one's end* (which is an act of inner freedom equivalent to the adoption of a maxim) and in so doing, one makes the law one's incentive, which, he says, affords the reward of moral pleasure associated with virtue. (MS 6:391) This application (maxim adoption) is also evident in the *Religion* where he says one 'should become not merely *legally* good but *morally* good . . . i.e. virtuous according to the intelligible character [of virtue] (*virtus noumenon*) and thus in need of no other incentive to recognize a duty except the representation of duty itself' (R 6:47). There are many references to the second application (i.e., of constraining oneself *to act* on a moral maxim) in the *Doctrine of Virtue* and elsewhere. For example, 'Virtue is the strength of the human being's maxims in fulfilling his duty' (MS 6:394). It is natural that virtue should be associated with both of these applications when we think of it as a developed inner freedom. One who adopts a moral maxim with relatively little interference from his sensuous side is also the sort of person who, from free habit, acts on this maxim consistently. We may even say that these 'two' applications of virtue might just be two aspects of the same application - at least in the case of duties of virtue which consist in setting an end and continually pursuing it - since if I consistently fail to *act* and thereby fail to pursue that end of virtue, it can hardly be said that I have *adopted* that maxim (i.e., set that end). It is virtue itself which allows us to adhere to our (long and difficult) task in pursuing wide duties of virtue.

Returning to the main issue at hand, having outlined the notion of virtue as a free habit, Kant tells us that inner freedom *requires*[133] 'being one's own *master* in a given case . . . and *ruling* oneself . . . that is, subduing one's affects and *governing* one's passions.' (MS 6:407) He then goes on to give a brief account of each of these. He tells us that an affect is a sudden feeling which, in preceding reflection, 'makes this [reflection] impossible or more difficult'. (MS 6:408) In the *Anthropology*, he says this feature of it prevents us from deciding whether to

---

[133] It may seem strange to claim that self-mastery and self-governance over one's affects and passions are *required for* inner freedom when, as we have seen, inner freedom just is a capacity for rational self-constraint. Although he does not highlight this unclarity in Kant's exposition, Engstrom's (clearer) view is that to possess inner freedom is *to be* independent of the influence of affects and passions; self-mastery and self-governance are components of inner freedom because through these one achieves said independence (2002, p.310).

give ourselves over to the affect (VA 7:251). It also seems then that in pre-empting reflection and in being rash, an affect also interferes with our ability to judge which specific action is required to satisfy it since he says, 'it makes itself incapable of pursuing its own end' (VA 7:253). These features make it clear that affects are inimical to the exercise of inner freedom.

In contrast, Kant says that a passion is 'a lasting inclination (e.g., hatred, as opposed to anger).' (MS 6:408) It permits calm reflection and thereby allows principles to be based upon it. Where the original inclination is contrary to the law, its deepening into a passion involves brooding and a premeditated incorporation into a maxim: it is then 'a true *vice*.' (MS 6:408)

In the *Anthropology* Kant regards *all* passions as evil since, as a kind of obsession, a passion may lead to the neglect of ends some of which may be required (VA 7:267). Allen Wood points out that passions are 'also opposed to prudence' (1999, p.252) because their influence is such that we fail to give our other desires and interests proportionate attention. Finally, passions are always directed to persons (VA 7:270). Once again, we can see why Kant thinks the exercise of inner freedom is being curtailed and thus why the agent must address his passions.

Although Engstrom does not mention this point (nor to my knowledge, does Kant), the thought seems to be that virtue can operate as both prevention and cure and that it takes on the former role as a promptitude in moral choice: in allowing us to choose a moral maxim of action more readily than we could without it, it thereby *prevents* the choice of an evil one in its place. However, it also has a curative role in cases in which we already have an evil maxim - say, a passion instead of a moral maxim. Let us examine how Engstrom thinks it can perform this role.

Kant seems to think that because self-mastery and self-governance oppose affects and passions which are inimical to inner freedom, that self-mastery and self-governance are *aspects* of inner freedom - at least this is how Engstrom takes it. We might wonder, then, how these two elements of inner freedom allow the will independence from influence by affects and passions. Kant does not say but Engstrom thinks that he can reconstruct an answer (2002, p.310). He says that self-mastery is negative and correcting and directed towards passions, whereas self-governance is positive and habilitating and addresses affects. He states that this picture is based on the distinction between discipline and culture in the first *Critique* (A709/B737-738) and also the notion that passions are associated with actual evil whereas affects are associated with mere lack of virtue (MS 6:408). He does not elaborate on how it is based on these distinctions. However, the thought seems to be that the negative, correcting approach (self-mastery) is suitable for something evil, and the

positive, and the habilitating one (self-governance) is appropriate for mere affects,[134] which Kant describes as 'something childish and weak' rather than malign (MS 6:408).

Engstrom's proposal is that in self-mastery, we limit inclinations generally and even extirpate those which directly oppose duty so that they cannot degenerate into passions. It is not clear that either contemplation (or anything else) can rid us of an inclination entirely (Kant says 'to want to extirpate them [natural inclinations] would not only be futile but harmful and blameworthy as well' (R 6:58)). Nevertheless, it may still be possible to prevent an inclination becoming the sort of gnawing obsession that is a passion by disciplining ourselves in the way we see the object of that passion through contemplation of the dignity of the law. Self-governance, on the other hand, Engstrom supposes, would consist in off-setting the disruptive effects of affects on judgement and understanding by cultivating these. He introduces Kant's dual approach for the acquisition of virtue from the *Doctrine of Virtue* - *contemplation* and *practice* - and thinks these can lend some detail to the workings of self-mastery and self-governance. Kant says,

> this capacity as *strength* (*robur*) is something that must be acquired, through the elevation of the moral *motive* [*Triebfeder*] (the representation of the law) through contemplation (*contemplatio*) of the dignity of the pure rational law in us, and at the same time also through practice (*exercitio*). (MS 6:397; quoted in Engstrom, 2002, p.311)

Engstrom explains how he thinks the first approach - *contemplation* of the dignity of the law in ourselves and others - may help us to *discipline* a passion. Such contemplation allows us to see more clearly that others are *persons* to be respected. This militates against the development of passions as these, he says, require that we misrepresent our relation with others as persons. The other method which Kant recommends in this passage - *practice* - Engstrom thinks allows us to *cultivate* our powers of pure practical reason, (as well as understanding and judgement) 'through which is achieved the self-composure wherein they are not liable to interference from feelings.' (2002, p.311) The idea seems to be that although the strong person may still experience an affect, he has the self-control not to act on it.

I think Engstrom has not said anything false here about discipline through contemplation but he omits what I take to be a crucial element in it: in contemplating the *dignity* of the law in myself (which is what Kant proposes in the quotation), I am contemplating what gives it, (and me), dignity; namely, the fact that as the law of a *free*, rational being, its dignity - and

---

[134] We might object that in the passage quoted above in which Kant mentions these two elements of inner freedom (MS 6:407) he says we must *govern* our *passions*. Nevertheless, this discrepency arguably does not invalidate Engstrom's subsequent concrete suggestions as to how one develops inner freedom as strength.

mine, for it is my law - lies in its ability to elevate me above the causality of nature. In contrast, the licentious conception of freedom associated with a passion would have me shackle myself to obsessive demands. In short, in contemplating this dignity, the incomparable value of *freedom* is foregrounded in my mind and so I strengthen my resolve to act *from inner freedom* (in this case, to dissolve a passion). I may also foreground the knowledge that the object of my passion is a *person*[135] in whom the law and its dignity of freedom also inheres and so I see I must respect her.

Incidentally, Engstrom seems to be in two minds as to whether he is discussing the *exercise* of virtue or the *acquisition* of it when explaining how contemplation may overcome passions and practice may cultivate affects. For example, he says contemplation is an activity *in which virtue consists* (2002, p.311) when commentating on the passage (MS 6:397) in which Kant introduces contemplation and in which Kant explicitly states that it is a method acquisition. In addition, as we shall see shortly, Engstrom continues his discussion of contemplation by going on to explain how it is a key technique for the *acquisition* of virtue and how this is shown in the Doctrine Of Method of the second *Critique*.[136] Perhaps the reason for this slight muddle, is that it might be that contemplation is both an *exercise* and a *method of acquiring* virtue since it consists in becoming conscious of the dignity of a law of freedom as it relates to a particular bad action and can aid me in overcoming that particular vice *now* (so in this sense contemplation is a form of the exercise of virtue). But this contemplation can also aid me in being virtuous more generally: i.e., so long as I am conscious of the law's dignity as a law of freedom, such contemplation will have enhanced my ability to overcome *any* obstacle to morality more generally.

4.4 The acquisition of virtue: contemplation, practice and the self-consciousness of freedom

Engstrom associates (I think incorrectly) the sort of contemplation Kant recommends in his key passage on the acquisition of virtue (MS 6:397), (where Kant says we must contemplate the dignity of the law) with the sort of contemplation we might encourage in the young - the contemplation of real-life examples of good deeds done from duty discussed in the Doctrine of Method of both the second *Critique* and the *Doctrine of Virtue*. I think contemplation of the dignity of the law and contemplation of examples are distinct strategies but that both aim to develop virtue through the consciousness of inner freedom in us. We saw above how the

---

[135] As we saw above, the object of a passion is always a person (VA 7:270).

[136] Although (confusingly), Engstrom does not seem to realise that the sort of contemplation described in the Doctrine of Method is different from that advocated at MS 6:397 (although it is also concerned with the acquisition of virtue). I examine the difference between these sorts of contemplation, below.

contemplation of the dignity of the law may issue in greater virtue by clarifying *the value of freedom* and strengthening my resolve to act as a free being. Contemplation of examples, on the other hand, I think works by inspiring pupils to emulate the sort of exercise of inner freedom which they see is *possible* in certain examples.

Whilst, Engstrom is wrong to think that contemplation of the dignity of the law is the same as contemplation of examples, his rendition of how the latter approach works is correct. Let us examine this now: he says, the idea in the use of examples is that seeing that pure reason alone can be practical inspires the pupil to believe (truly) that he is capable of comparable deeds. Rather than being models to be imitated, they ought to be 'proofs of the practicability (*Tunlichkeit*) of free self-determination'. These are thought to 'inspire the freedom that is in us by making manifest its exercise in another.' (Engstrom , 2002, p.312). Engstrom thinks that the way the examples are supposed to work shows that Kant takes the acquisition of inner freedom as strength fundamentally to involve what might be described as *freedom's attainment of self-consciousness*. (ibid.)

If we analyse one of the key passages in the Doctrine of Method of the second *Critique*, it is evident that he is right in this. In this discussion of the use of examples (KpV 5:160-161), the pupil is presented with examples of actions carried out by persons with developed inner freedom: i.e., virtue: these are examples in which 'no incentives of inclination have any *influence* on it [the will] as determining grounds' (KpV 5:160; emphasis added). Kant says, 'the pupil's attention is fixed on the consciousness of his *freedom*' (KpV 5:160) This freedom is his capacity for inner freedom. Kant says this explicitly in the following claim: 'there is revealed to human being an inner capacity not otherwise correctly known to himself, the *inner freedom* to release himself from the impetuous importunity of inclinations' (KpV 5:161). The thought seems to be that with sufficient exposure, eventually inclinations 'not even the dearest, has any influence on a resolution' (KpV 5:161). Our pain at the renunciation of them is replaced by a feeling of respect for ourselves such that where previously we might have been concerned over the denial of pleasure, we now *dread* the self-contempt we would feel over wrong-doing. This is 'the *respect for ourselves* in the consciousness of our freedom.' (KpV 5:161) Contemplation is not restricted to exposure to examples of other people's virtue. It can also focus on one's own actions. Given this, it seems reasonable to suppose that one may also be aware of one's capacity for inner freedom whilst *exercising* that freedom.

5. Virtue underpins all other core elements of moral development

At the beginning of Section 3, the major elements of moral development were said to be uprooting evil maxims, the duty of moral perfection, the duty to cultivate those empirical motives which can co-operate with duty (especially compassion), the duty to be holy, the need for reassurance of one's goodness and the duty to acquire virtue. In establishing how virtue can be acquired in the previous section, I have completed the account of the possibility and method of pursuit of these, the major elements of moral development. The next task is to show how virtue (*1*) aids the *pursuit* of these elements the various ways and (*2*) part-constitutes their *intended end-states*. Regarding (*1*), the sorts of difficulty in pursuing ends of moral development which can be remedied by virtue and which I wish to examine are firstly, the *arduousness* of the pursuit of at least some; secondly, possible *anxiety* over revelations about one's personal worth; thirdly, the possible feeling of *dejection* over a literally interminable set of tasks. Let us turn to (*1*).

Duties in general are not done with the aim of any kind of gratification and so (depending on the duty and the level of virtue of the agent), even if not otherwise unpleasant to do, they may still be a chore, and duties of moral development are no different in this regard. We can see this, for example, in the pursuit of the duty of moral perfection - the duty to do all of one's duties. To pursue this successfully, one must be attentive to the necessitating thought of the law in *all situations* since many situations may generate a required action and it may be hard to predict which ones will. Similarly, as we have seen repeatedly, many elements of moral development require self-examination and this again requires effort, probably without any attendant pathological pleasure. For example, whilst many evil maxims left over from before the revolution may be revealed (perhaps defeasibly) simply as the situations to which they pertain come up - i.e., without the agent having to do anything for them to be revealed - we noted that the good agent as such would wish to take a proactive approach and also root out such maxims himself. He might do this by reflecting on past behaviour in the various sorts of situation which occur repeatedly in his life (but have not occurred since the revolution): his disparaging treatment of the irritating colleague he sees most days, the street people he passes without a glance whilst commuting and so on. It can be tempting not to do work that is hard and does not even aim at pleasure or material benefit, especially when that effort aims at eliminating what were seen as liberating exercises of outer freedom. However, such arduousness is perhaps the least of the difficulties facing the person engaging in moral development.

What is perhaps more challenging is that the self-examination involved in many of the elements of moral development can be an *anxious* affair. Kant certainly thinks it is daunting when he says, 'Only the descent into the hell of self-cognition can pave the way to godliness.'

(MS 6:441) Perhaps the most obvious example of this is in the revelation of outright evil maxims. This is likely to be an unpleasant and difficult thing to accept, especially at the beginning of the development process, since the new recruit now with aspirations to be good will find no joy in having to expose many self-deceptive narratives for what they are, thereby foregoing their comfort and protection for the first time and facing up to the extent of his evil. He will also, in all probability, have to confront his having been one who not only violated rights but compounded the offense by doing so with the pompous attitude of a supposed entitlement to do so.

In addition, actions need not be potentially evil to give the agent reason to be anxious about what he will find. The necessity to examine motives for actions *in accordance with duty* is something about which an agent might be apprehensive since a great deal can be at stake. For example, the volunteer at a refuge for the homeless, whose supposed unselfish devotion over a large portion of his life *had* been greatly responsible for sustaining his sense of self-worth in that time, who has discovered that it was all probably done entirely for his own gratification could very well find such a revelation devastating.

In addition, we have seen that whilst an agent may glean information about her interests just by reflecting on the way she sees a situation during the process of practical deliberation, it may be possible to glean more detailed information but only by persisting in asking questions beyond the point of deciding what to do and this may be difficult: if by the time she has reached this point (of deciding her action), she has not discovered anything untoward about her interests but is aware that by persisting, she may discover something which is difficult to accept, there may be a great temptation to stop deliberating especially since choosing an end is a natural cut-off point for practical deliberation. It may take a special effort and a resilience towards unpleasant findings to persist past the point of choice in this manner.

In fact, given that it is the new recruit to morality, who is more likely than others to discover a relatively large number of evil and impure maxims (since he has had less opportunity than them to deal with these) and given that he is not yet used to being without the protection of self-deception and further that he has a relatively undeveloped virtue, it may be necessary for him to pursue ends such as gleaning information on his motives (and rooting out evil maxims) with perhaps less vigour than he might otherwise would wish until he has acquired the necessary virtue to continue. If overwhelmed, he may revert to an overarching policy of evil. Unfortunately, we must suppose that not every moral beginner will realise this and may revert back to evil. In addition, telling himself he needs a break is an ideal self-deceptive narrative to facilitate a fall.

The third problem area concerns the interminability of the task and the agent's morale. Duties of moral development are wide because the ends involved (e.g., virtue, holiness) cannot be achieved in a lifetime (nor even in the afterlife, Kant thinks). I submit that for some people at least, the pursuit of a task which has no definite, reachable ending is for that reason apt to be disheartening. Worse still, if the task is such that one cannot have the satisfaction of finishing (because it has no ending), one might at least hope for reliable signs that one is progressing. But, as we have noted, whilst such indications given by the methods we have considered are fairly reliable in the early stages (when most of the news concerning our moral condition is bad), they are far less reliable when the news is apparently good (because of the possibility of self-deception). This means at the latter stages of moral development, where presumably the news would mostly appear to be good to the agent (and unknown to him must actually *be* good or else he would not be at the latter stages), the agent will constantly have to treat this information with extreme caution. This may also be disheartening. Fortunately, one at the latter stages should have a level of virtue better able to cope with such uncertainty. Moreover, if virtue can be the joyous affair which (we discovered in Sub-section 3.4) Kant thinks it is, this may off-set, to some extent, any hardship felt.

In addition to our requiring virtue to combat the difficulties associated with the pursuit the elements of moral development, it also part-constitutes the intended end-states of many (perhaps all) of these elements. For example, the extent to which I succeed in pursuing the end of moral perfection will vary according to how morally strong I am, *ceteris paribus*. Quite simply, if I am stronger, I can more readily do my more challenging duties and so I am better able to do more of my duties. Similarly, since the duty to be holy commands me to purify those maxims which have mixed grounds, if I am to continue to hold those maxims from duty alone and act on them from it alone my virtue may need to be greater than when I relied upon co-operating empirical motives. In addition, if one does not purify but instead regulates the empirical motives in such maxims, again, it presumably requires strength to constrain the empirical desires involved.

Virtue seems to have a wide variety of applications: Engstrom has explained how as a free habit, it allows the agent to be prompt in choosing the right maxim from duty (or in doing the right thing); he has also highlighted Kant's emphasis on its use in the cultivation of affects and the mastery of passions; I have argued that it is needed to overcome the arduousness, anxiety, and dauntingness of the pursuit of the elements of moral development and that it part-constitutes the aimed for end-states of duties such as the duty to be holy. However, we must not suppose from this wide range of applications that virtue is something different in each case. Whilst it seems necessary to know what the specific problem is in order to bring virtue to bear upon it - for example I need to have some indication that a certain empirical

motive requires cultivation before I can cultivate it - virtue nevertheless always consists in the consciousness of one's inner freedom as a capability or strength to do what the law (freedom) demands. The agent in becoming inspired to do as the law demands by recognizing its incomparable value as true freedom and by acknowledging in himself his ability - his inner freedom - to do what it demands, may complete any task that the moral project demands of him.

Conclusion

This story began with the idea that the free will - freedom itself - may freely shackle itself: we saw that in choosing to take the licentious pursuit of outer freedom as its conception of freedom *simpliciter* and in choosing to maintain this deception, the will adopts an evil *Denkungsart* through which it restricts itself to worldly pursuits which follow the paths of natural necessity. In doing this, it is further constrained by a false table of values - valuing its ability to express outer freedom despite other wills - a way of valuing which conceals what is truly valuable and incomparably so: the moral law and persons in whom that law inheres (since these are respectively, the *expression* and the *locus* of the true freedom of autonomy).

However, the will may also choose to throw off these shackles in a process *initiated* both in the rejection of the evil *Denkungsart* (despite its mutually supportive elements) and in the embrace of autonomy as true freedom: in the sincere resolution to be good. This is, however, only an initial step because, as we saw, Kant believes that the good will as such affirms its commitment in the form of moral progress. When expounding the revolution earlier, I said that this notion - that affirmation follows a sincere resolution in a good person - could also be explained in terms of freedom's determination to free itself *fully*. Now that we have the account of autonomy and of virtue as two forms of inner freedom, I am in a position to explain what I meant by that. In becoming conscious of its autonomy, the will becomes aware of itself as inner freedom as sheer capacity. But it strikes me that the will *qua* freedom is only satisfied with the fullest expression of freedom that it can manage (whatever its conception of freedom might be) and so the *good* will, conscious that freedom is autonomy - i.e., inner freedom as capacity - is only satisfied with the fullest expression of inner freedom that it can manage (just as the evil will was only satisfied with the fullest expression of outer freedom *it* could manage). The way the good will as such strives to enhance and extend its exercise of inner freedom is in the pursuit of the various elements of moral development: in striving to eradicate all evil maxims, in doing all of its duties and from duty, in ensuring that those empirical motives which are tolerated at least do not enslave it and in developing its virtue.

Since both the pursuit of these elements and their intended end-states all depend to varying degrees on the will's virtue and since this is acquired by being conscious of its ability to exercise inner freedom, it turns out that to continue the process of freeing itself through the elements of moral development, the will must become conscious not only of its inner freedom as capacity (its autonomy) at the initial stage of development (the resolution) but also of its inner freedom as enhanced capability in the on-going moral project. The good *Denkungsart* is distinguished from the evil one not just as an attitude to choice, that is, by its striving for maximal inner freedom in its choices rather than maximal outer freedom. It is also distinguished by its seeking self-consciousness instead of self-deception, by being vigilant about the situations in which self-deception is a threat (for example in the pursuit of moral self-knowledge) and aware of the way self-deception works (for example its exploitation of similarities between the true and the false) rather than being negligent in these matters. Of the will which strives to do all of this to the best of its ability, we can say that freedom has, through self-awareness, striven to make itself truly free.

Morgan (2005, p.106-107) takes his rational reconstruction of evil to show that Kant's claims about the propensity to evil are more than just an *ad hoc* afterthought but are central to his moral philosophy and that pessimism is therefore integral to his ethical thought. With the emergence of the formal proof of evil, Morgan argues, contemporary Kantians may have to reflect this pessimism in their neo-Kantian moral psychology. This would make the practical philosophy perhaps a less congenial resource to 'contemporary post-Enlightenment liberal optimism' than was supposed. And those chapters of the present study which develop the notion of an evil *Denkungsart* would seem only to make matters worse since they throw into sharper relief the arrogant and wilful refusal of evil to acknowledge itself for what it is. I hope, however, that this study, in highlighting those ways in which such an attitude may be overcome despite its recalcitrance, in indicating the ways in which the Kantian agent can improve her prospects for continued development following the revolution despite the opacity of maxims and the persistent threat of self-deception and the constant lure of the incentive to license - that all this might mitigate that pessimism which Morgan thinks Kant-inspired ethical thought may have to inherit. The possibility, shown here, that the human will can be unified in this way, making the agent more willing to do her duty, may go some way to help dispense with the caricatured view of the Kantian moral life as one limited to struggle and hardship, as some virtue theorists might have us believe. Kant does not avoid the fact that moral life can be difficult. But what the would-be virtuous individual found hard can, with effort, become easier and sometimes even joyous since it is performed by a rational being no longer at war with the world and himself.

References

Allison, H. (1990). *Kant's Theory of Freedom*. Cambridge: CUP.

Beck, L.W. (1960). *A Commentary on Kant's* Critique of Practical Reason. Chicago: The University of Chicago Press.

Bach, K. (1981). An Analysis of Self-Deception. *Philosophy and Phenomenonological Research, 41* (3), 351-370.

Bach, K. (1985). More on Self-Deception: Reply to Hellman. *Philosophy and Phenomenonological Research, 45* (4), 611-614.

Baxley, A. M. (2003). Autocracy and Autonomy. *Kant Studien, 94* (1), 1-23.

Benson, P. (1987). Moral Worth. *Philosophical Studies, 51* (3), 365-382.

Caswell, M. (2006a). Kant's Conception of the Highest Good, the *Gesinnung*, and the Theory of Radical Evil. *Kant-Studien, 97* (2), 184-209.

Caswell, M. (2006b). The Value of Humanity and Kant's Conception of Evil. *Journal of the History of Philosophy, 44* (4), 635-663.

Cicero, M.T. (2000). *On Obligations*. (P.G. Walsh Trans.). Oxford: OUP.

Crisp, R., & Slote, M. (Eds.). (1997). *Virtue Ethics* . New York: OUP.

Engstrom, S. (2002). The Inner Freedom of Virtue. In M. Timmons (Ed.). *Kant's* Metaphysics of Morals*: Interpretative Essays*. New York: OUP.

Green, M.K. (1992). Kant and Moral Self-Deception. *Kant-Studien, 83* (2), 149-169.

Herman, B. (1993). *The Practice of Moral Judgment*. Cambridge, MA.: Harvard University Press.

Henson, R. (1979). What Kant Might Have Said: Moral Worth and the Overdetermination of Dutiful Action. *The Philosophical Review, 88* (1), 39-54.

Kant, I. (1899). *Kant on Education (Uber Pädagogik)*. A. Churton (Trans.). London: Kegan Paul, Trench, Trubner & Co.

Kant, I. (1983). Speculative Beginning of Human History. In T. Humphrey (Trans.). *Perpetual Peace and Other Essays*. Indianapolis: Hackett Publishing Co.

Kant, I. (1996). *Critique of Practical Reason*. In M. Gregor (Ed. & Trans.). *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (pp. 133-271). New York: CUP.

Kant, I. (1996). *Groundwork of the Metaphysics of Morals*. In M. Gregor (Ed. & Trans.). *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (pp. 37-108). New York: CUP.

Kant, I. (1996). *Religion Within the Bounds of Mere Reason*. In A. W. Wood (Ed.) (G. di Giovanni, Trans.). *The Cambridge Edition of the Works of Immanuel Kant: Religion and Rational Theology* (pp. 39-215). New York: CUP.

Kant, I. (1996). *The Metaphysics of Morals*. In M. Gregor (Ed. & Trans.). *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (pp. 353-603). New York: CUP.

Kant, I. (1998). *Critique of Pure Reason*. (P. Guyer & A. Wood Ed. & Trans.). New York: CUP.

Kant, I. (2000). *Critique of the Power of Judgment*. (P. Guyer Ed. & Trans.) (E. Matthews Trans.). New York: CUP.

Kant, I. (2006). *Anthropology from a Pragmatic Point of View*. R.B. Louden (Ed. & Trans.) New York: CUP.

Korsgaard, C.M. (1996a). *Creating the Kingdom of Ends*. Cambridge: CUP.

Korsgaard, C.M. (1996b). *The Sources of Normativity*. Cambridge: CUP.

MacIntyre, A. (1981). *After Virtue: A Study in Moral Theory*. London: Duckworth.

Morgan, S. (2005). The Missing Formal Proof of Humanity's Radical Evil in Kant's *Religion*. *The Philosophical Review, 114* (1), 63-114.

Morgan, S. (2006). Kant on Self-Conceit. Unpublished book chapter retrieved from http://www.bristol.ac.uk/philosophy/department/staff/sm.html

Munzel, G.F. (1999). *Kant's Conception of Moral Character: The "Critical" Link of Morality, Anthropology, and Reflective Judgment*. Chicago: The University of Chicago Press.

Nietzsche, F. (1996). *On the Genealogy of Morals*. Oxford: OUP.

O'Hagan, E. (2009). Moral Self-Knowledge in Kantian Ethics. Ethical Theory and Moral Practice, *12* (5), 525-537.

O'Neill, O. (1989). *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: CUP.

Pasternack, L. (1999). Can Self-Deception Explain *Akrasia* in Kant's Theory of Moral Agency? The Journal of The Southwestern Philosophical Society, *15* (1), 87-97.

Reath, A. (2006). *Agency & Autonomy in Kant's Moral Theory*. Oxford: Clarendon Press-OUP

Sartre, J.-P. (1957). *Being and Nothingness: an Essay on Phenomenonological Ontology*. (H.E. Barnes Trans.). London: Methuen.

Sidgwick, H. (1962). *The Methods of Ethics* (7th Ed.). London: MacMillan & Co. Ltd.

Sherman, N. (1993). Wise Maxims/Wise Judging. *Monist, 76* (1), 41-65.

Sherman, N. (1997). *Making a Necessity of Virtue: Aristotle and Kant on Virtue*. Cambridge: CUP.

Sussman, D. (2005). Perversity of the Heart. *The Philosophical Review, 114* (2), 153-177.

Webber, J. (2009). *The Existentialism of Jean-Paul Sartre*. London: Routledge.

Williams, B. (1993). *Ethics and the Limits of Philosophy*. London: Fontana Press.

Wolff, R. (1986). *The Autonomy of Reason*. Gloucester, MA: Peter Smith.

Wood, A.W. (1984). Kant's Compatibilism. In A. W. Wood (Ed.). *Self and Nature in Kant's Philosophy*. London: Cornell University Press.

Wood, A.W. (1999). *Kant's Ethical Thought*. New York: CUP.